U.S. DEPARTMENT OF COMMERCE

PB80-168065

Approaches to Analyzing Data from the Portland Study, Task 4

Technology Service Corp.
Santa Monica, CA

Prepared for

Environmental Protection Agency
Ann Arbor, MI

May 79

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.
TSC-PD-A158-5

2.

3. RI PB80-168065

4. TITLE AND SUBTITLE
Approaches to Analyzing Data from the Portland Study, Task 4

5. REPORT DATE
May 1979

6. PERFORMING ORGANIZATION CODE

7. AUTHOR(S)
Larry S.L. Lai
John D. Gins

8. PERFORMING ORGANIZATION REPORT NO.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Technology Service Corporation
2811 Wilshire Blvd.
Santa Monica, California 90403

10. PROGRAM ELEMENT NO.

11. CONTRACT/GRANT NO.
68-03-2392

12. SPONSORING AGENCY NAME AND ADDRESS
U.S. Environmental Protection Agency
2565 Plymouth Rd.
Ann Arbor , Michigan 48105

13. TYPE OF REPORT AND PERIOD COVERED
Final Report

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The EPA is currently conducting the Short Test Correlation and Effectiveness Study (Portland Study), a large-scale emission testing program, in Portland, Oregon. For all the vehicles involved in the study, vehicle emissions are measured by three short tests (STs) at a state inspection station and by the three tests plus the Federal Emission Certification Test Procedure (FTP) at an emission testing laboratory.

The two basic areas of interest in the Portland Study are: 1.) the issue of association of short-test emissions (as measured in a real-world inspection/maintenance environment) and FTP emissions; and 2.) theoretical and practical questions concerning the process of inspection and maintneance.

The purpose of this report is to provide various sound statistical procedures that could be used in analyzing data from the Portland Study. In particular, the report is written in the form of a users' manual, with two types of readers in mind: 1.) The reader involved in the Portland Study who has a limited knowledge of statistics. 2.) The reader who is an experienced programmer.

KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Motor Vehicles Emission Exhaust Gases Tests Inspection Correlation Air Pollution	Portland Study Inspection/Maintenance Programs and Short Test Correlation and Effectiveness Study	

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

DISTRIBUTION STATEMENT ELEASE UNLIMITED	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21. NO. OF PAGES
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE

Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

Technology Service Corporation

2811 WILSHIRE BOULEVARD • SANTA MONICA, CALIFORNIA 90403 • PH. (213) 829-7411

1.

TSC-PD-A158-5

FINAL REPORT

APPROACHES TO ANALYZING DATA FROM THE PORTLAND STUDY

TASK 4

Contract No. 68-03-2392

May 1979

by

Larry S.L. Lai

John D. Gins

Prepared for:

Ms. Janet Becker
Characterization and Application Branch
Office of Air and Water Programs

Submitted to:

U.S. Environmental Protection Agency
2929 Plymouth Road
Ann Arbor, Michigan 48105

ACKNOWLEDGMENT

During the preparation of this report, valuable guidance and assistance were provided by Ms. Janet Becker of the Environmental Protection Agency's Office of Air and Water Programs, who served as EPA Project Officer for the Portland Study.

Mr. John Gins was the Principal Investigator for the project, and Dr. Larry Lai, the principal author. Dr. Leo Breiman, Senior Consultant, Dr. John Eldon, Senior Scientist, and Ms. Nancy Chang, Research Statistician, made numerous valuable suggestions and comments on the early drafts of the report. Ms. Marian Branch reviewed and assembled the many drafts of the manuscript.

CONTENTS

1.0 INTRODUCTION 6

1.1 BACKGROUND 6

1.2 GENERAL DESCRIPTION 6

1.3 USER'S GUIDE 8

2.0 SHORT TEST/FTP ASSOCIATION 10

2.1 PROCEDURES FOR ANALYZING THE ST/FTP ASSOCIATION 18

2.1.1 Stage 1. Graphical Exploration 20

2.1.2 Stage 2. Assessment of the Distributional Properties of the Emission Data 23

2.1.3 Stage 3. Implementation of Classification Algorithms 45

2.1.4 Stage 4. Evaluation of the ST/FTP Association . . 71

2.1.5 Stage 5. Testing Hypotheses 78

3.0 RESPONSES TO SPECIFIC QUESTIONS 101

3.1 RESPONSES TO SPECIFIC QUESTIONS RELATED TO THE ST/FTP ASSOCIATION 102

3.2 RESPONSES TO QUESTIONS RELATED TO IMMEDIATE EFFECTIVENESS OF AN INSPECTION/MAINTENANCE PROGRAM . . . 112

3.3 RESPONSES TO QUESTIONS RELATED TO LONG-TERM EFFECTIVENESS OF INSPECTION/MAINTENANCE 128

APPENDIX A. STATISTICAL TABLES 142

APPENDIX B. MASTER-AID CLASSIFICATION ALGORITHM USER'S MANUAL . . . 153

APPENDIX C. AS-51 LOG-LINEAR MODEL FIT ALGORITHM 169

REFERENCES 178

FIGURES

Number Page

2.1 Example of ideal situation in which short-test results determine FTP categorization 11

2.2 Example of common situation in which two point sets overlap 13

2.3a Example of short-test results with poor correlation, E*, with FTP 15

2.3b Example of desired probability of the misclassification computed with the formula $E^* = (1-P)E_0+PE_1$ 16

2.3c Example of case where one might need a nonparametric technique for generating a nonlinear rank score 17

2.4 Series of block-stages for ST/FTP association problem . . 19

2.5 Examples of histogram plots 21

2.6 Scatterplot, with points from Categories I and II identified by distinct letters 24

2.7a Simulated normal data plotted on normal paper 27

2.7b Simulated non-normal data plotted on normal paper 28

2.8 Linear regression line fit for assessing the joint normality of NO_x and CO 35

2.9a Distance representation of bivariate normal data on a semilog probability plot 43

2.9b Angle representation of bivariate normal data on a uniform probability plot 44

2.10 Hypothetical decision tree 62

2.11 Tree structure of the Output Example 68

TABLES

Number Page

2.1 Bivariate HC and NOX Values with Associated Z_1, Z_2, θ, r^2 and $r^2 \log_e r^2$ 39

2.2 Sorted Values of $U = 0.5 + \frac{\theta_k}{2\pi}$, the Associated Cumulative Frequency $F = \frac{k}{n+1}$, and the KS Statistic 41

2.3 Output Example (Friedman's Algorithm) 65

2.4 Complete Models in Three Dimensions 87

1.0 INTRODUCTION

1.1 BACKGROUND

Under EPA Contract No. 68-03-2513, the Environmental Protection Agency is currently conducting the Short Test Correlation and Effectiveness Study (Portland Study), a large-scale emission testing program, in Portland, Oregon. For all the vehicles involved in the study, vehicle emissions are measured by three short tests (STs) at a State inspection station and by the three tests plus the Federal Emission Certification Test Procedure (FTP) at an emission testing laboratory.

The two basic areas of interest in the Portland Study are: (1) the issue of association of short-test emissions (as measured in a real-world inspection/maintenance environment) and FTP emissions; and (2) theoretical and practical questions concerning the process of inspection and maintenance.

1.2 GENERAL DESCRIPTION

The purpose of this report is to provide various sound statistical procedures that could be used in analyzing data from the Portland Study. In particular, the report is written in the form of a user's manual, with two types of reader in mind:

1. The reader involved in the Portland Study who has a limited knowledge of statistics. He or she should be able to follow the text, build up his/her own statistical capability with regard to the Portland Study, and make sound statistical judgments.
2. The reader who is an experienced programmer. He or she can easily implement the procedures, following the instructions and programs contained in the report.

Because of the many possible salient features of the data collected from the Portland Study, the report provides the user with a sequential approach to data analysis. Typically, the approach consists of a series of block-stages. Several subtasks are to be performed within each block-stage, and the exact content of the subtasks depends on the outcome of the analysis in the previous stages. For the performance of each subtask, several computational procedures are presented. Whenever the computation can be done with a hand calculator, the procedure will be displayed step by step, illustrated with numerical examples to make the computations more transparent. Whenever the computation calls for a packaged statistical subroutine, the source of the program (in some cases, even the details of the program), the essentials of card preparation, the typical computer printout, and the interpretation of the printout, as well as the kind of statistical inference, are provided in great detail. Although the user could do a good analysis by blindly following the complete cycle of block-stages, human judgment and insight are also required in performing the analysis because data analysis is an art as well as a science.

The organization of the report is as follows: Chapter 2 deals exclusively with the main issue of the association between short-test emissions and FTP emissions. Following a description of the general philosophy of the method for analyzing the short-test association, the sequential approach to the analysis is presented by dividing Section 2.1 into five subsections. Section 2.1.1 describes various graphical and descriptive techniques for understanding the data and for suggesting possible ways of analyzing the data. Section 2.1.2 describes various procedures for testing the distributional properties of the emission data. The outcome will guide the next phase of the analysis. Section 2.1.3 describes the technique of multivariate classification. This technique allows the user to extract the maximum amount of information out of the multivariate data in evaluating the ST/FTP association. Section 2.1.4 describes various methods for evaluating the ST/FTP association. Section 2.1.5 describes ways of detecting those factors (e.g., engine type, model year) that affect the ST/FTP association.

Chapter 3 gives a detailed account of the responses to the theoretical and practical questions concerning the ST/FTP association and the process of inspection and maintenance. For each question, the answer contains either a detailed computational procedure or a reference to procedures previously mentioned.

Appendix A gives all the needed statistical tables; Appendix B, a user's manual and program description of Friedman's tree-structured algorithm; and Appendix C, the iterative proportional fit algorithm to be used in Section 2.5, "Stage 5. Testing Hypotheses."

1.3 USER'S GUIDE

For the evaluation of the ST/FTP association, the user should first read Section 2.0 to get a general idea of the sequential approach. If the user wants to get a quick answer to a specific question related to the ST/FTP association, he or she should read Section 3.1, "Responses to

Specific Questions related to the ST/FTP Association." For actually doing the analysis, the user should read the description in Section 2.1.5, "Testing Hypotheses," in which guidelines are given on which part of the data will be analyzed, and then proceed to the analysis according to Section 2.1 to Section 2.5, sequentially.

As to the process of inspection and maintenance, the user should consult Sections 3.2 and 3.3, in which specific answers to specific questions are given.

2.0 SHORT TEST/FTP ASSOCIATION

The following paragraphs set forth the general philosophy of the method for analyzing the short-test/FTP association.

Each car participating in the Portland Study undergoes a series of short tests at a State inspection station: the idle test, the Federal Three Mode, and the Federal Short Cycle. These short tests are repeated at the emission control laboratory, and a Federal Emission Certification Test Procedure (FTP) is also performed. Each car has sets of measured HC, CO, and/or NO_x values obtained from the short tests administered by State personnel at the State inspection station. For the idle test, an indication of whether the vehicle passed or failed a given set of idle HC and CO standards is given. For the FTP, an indication of passage or failure is also provided.

With this background in mind, we designate the aggregate of cars passing the FTP as Category I, and those failing the FTP, Category II. And, conceptually, let us assume that an observation of HC, CO, and NO_x values can be thought of as a single point in a three-dimensional space. In the event that the short-test (ST) results agree perfectly with the FTP results, there should be a clear separation between the totality of points associated with cars in Category I and the totality of points associated with cars in Category II. And, therefore, solely on the basis of its short-test results, a car could be classified unambiguously into either Category I or II. Figure 2.1 illustrates such an ideal situation, with the short-test results on HC and NO_x completely determining the FTP categorization. In this idealized situation, the line $a(NO_x)+b(HC)=C$

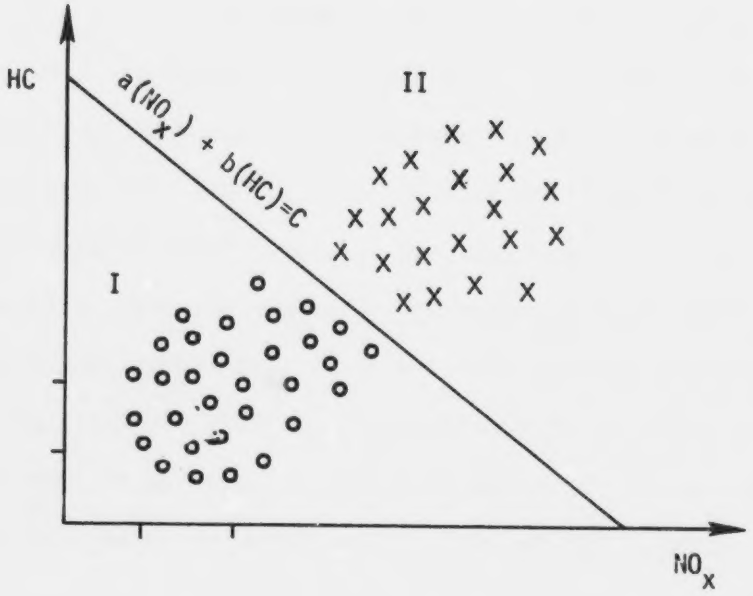

HC
$a(NO_x) + b(HC)=C$
II
I
NO_x

Figure 2.1 Example of ideal situation in which short-test results determine FTP categorization. The term C is the cutoff point for the linear rank score $a(NO_x) + b(HC)=C$.

would separate the two categories and provide a convenient way of establishing FTP passage or failure. For a given car, then, computing $a(NO_x)+b(HC)=C^*$, its short-test linear rank score, would classify the car into the "pass" category if $C^*<C$ and into the "fail" category otherwise.

Figure 2.2 illustrates the more common situation in which the two point sets overlap and a straight line will not separate the two sets completely, a situation, in fact, where the linear rank score corresponding to a straight line will produce two kinds of errors in classification, false positive error and false negative error. A false positive error means passage of the short test and failure of the FTP. A false negative error means failure of the short test and passage of the FTP.

To find the line which achieves a minimum of the sum of the positive and negative errors, we proceed as follows: Let E_0 denote the probability of producing a false positive error and let E_1 denote the probability of producing a false negative error. Assume that equal importance or cost is associated with errors of each type. According to a theorem in mathematical statistics, if the two batches of points are normally distributed, there always exists a straight line which separates the two sets with minimal expected probability of misclassification, E^*, which is the arithmetic average of E_0 and E_1, or the total number of misclassified points divided by the total number of points, or $(E_0+E_1)/2$ [Anderson 1958, p. 134]. The linear rank score corresponding to the line is called the "Fisher Linear Rank Score." Therefore, a Fisher Linear Rank Score will classify cars into Category I or II with minimal probability of

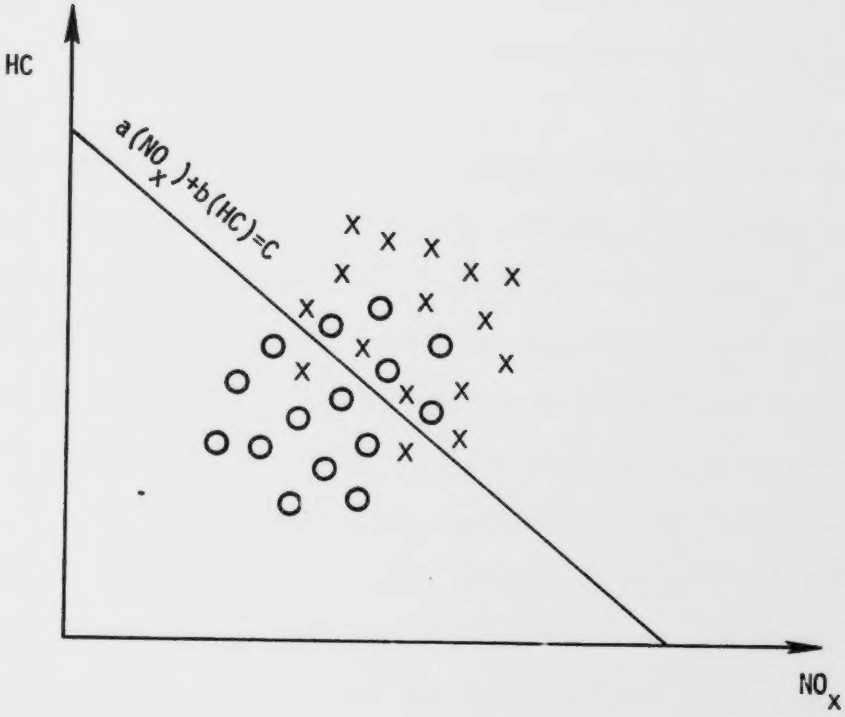


Figure 2.2 Example of common situation in which two point sets overlap. The minimal expected probability of misclassification, $(E_0+E_1)/2$, is as follows:

$$E_0 \approx \frac{2}{15}$$

$$E_1 \approx \frac{5}{15}$$

$$(E_0 + E_1)/2 = \left(\frac{1}{2} \times \frac{2}{15}\right) + \left(\frac{1}{2} \times \frac{5}{15}\right) = \frac{7}{30} \approx .23.$$

misclassification. E* can be used as a measure of the degree of correlation between the short test and FTP. Figure 2.3a illustrates a situation in which the short-test results are poorly associated with the FTP, based on HC and NO_x measurements (E*=.4).

The preceding discussions can be generalized to situations in which the observations are three-dimensional and each observation comes from Category I with probability P and from Category II with probability 1-P (i.e., the size of Category I may be P times the total number of cars in the Portland Study). Thus, a hyperplane may separate the two sets in space, and, accordingly, a linear rank score for the form $a(NO_x)+b(CO)+c(HC)$ can be used to perform the classification. The probability of misclassification can be computed with the formula $E^*=(1-P)E_0+PE_1$. (See Figure 2.3b.) For equal-sized groups, P=.5, and $E^*=.5(E_0+E_1)$, as before.

For real data, however, the two batches of points considered before may not follow normal distributions, and, because of large errors, a linear rank score thus becomes inappropriate in the process of classification. In this situation one may invoke a nonparametric technique to generate a "nonlinear" rank score to perform the classification. Figure 2.3c illustrates one such situation.

In view of the many possible salient features of the data, a sequential approach to the analysis is recommended. For this problem, the approach consists of a series of block-stages (see Figure 2.4). Within each block-stage, a number of trial runs will be performed to achieve a specific goal for each subtask within the block-stage. The

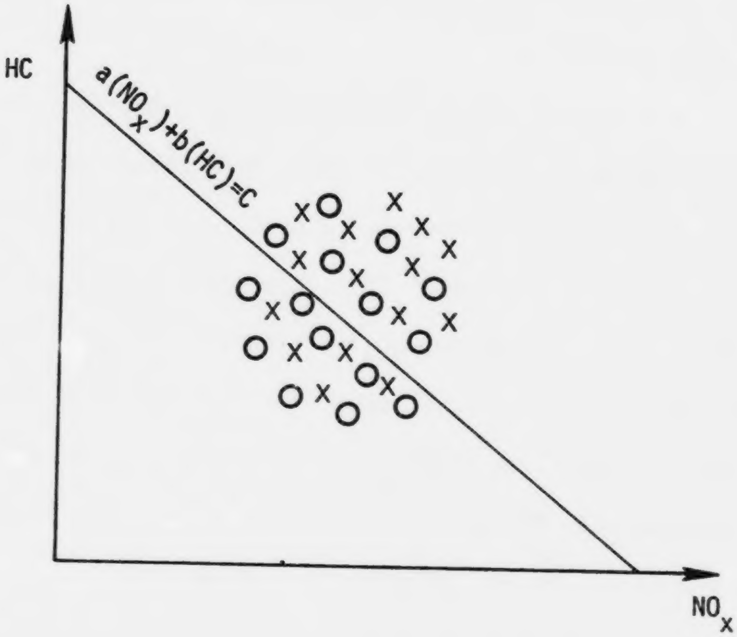


Figure 2.3a Example of short-test results with poor correlation, E^*, with FTP.

$$E_0 = \frac{7}{15}$$

$$E_1 = \frac{5}{15}$$

$$E^* = \left(\frac{1}{2} \times \frac{7}{15}\right) + \left(\frac{1}{2} \times \frac{5}{15}\right) = .4$$

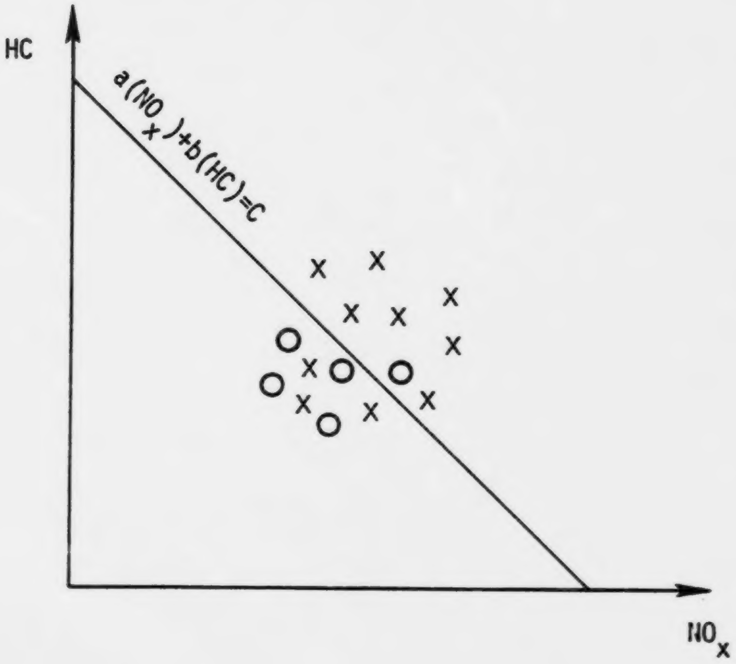


Figure 2.3b Example of desired probability of misclassification computed with the formula $E^* = (1-P)E_0 + PE_1$.

$$P = \frac{2}{3}$$

$$E_0 = \frac{1}{5}$$

$$E_1 = \frac{3}{10}$$

$$E^* = \left(\frac{1}{3} \times \frac{1}{5}\right) + \left(\frac{2}{3} \times \frac{3}{10}\right) = 0.27.$$

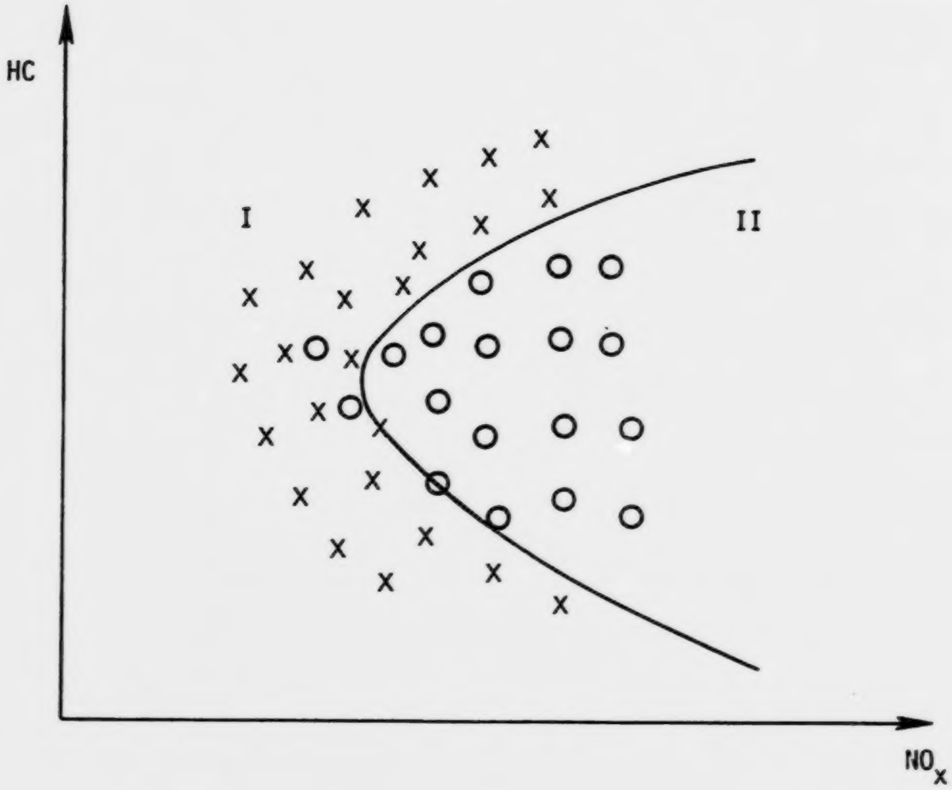


Figure 2.3c Example of case where one might need a nonparametric technique for generating a nonlinear rank score.

exact context of the subtask depends on the outcome of the analysis in the preceding stages. A feedback loop may be necessary between the two consecutive stages until a satisfactory answer has been reached.

2.1 PROCEDURES FOR ANALYZING THE ST/FTP ASSOCIATION

In this section, a sequential approach to analyzing the short-test/FTP association will be described. The input to the procedure consists of the at least five-dimensional observation (two, three, or nine for the ST measurements, three for the FTP measurements) on each vehicle. For a given set of FTP standards, the FTP "pass" or "fail" status of a vehicle can be determined through comparison of the three FTP emission values with the three standards, the violation of any one emission standard resulting in a "fail." As before, let I and II denote the sets of all cars which pass and fail the FTP, respectively. The following stages for analyzing the problem of association are recommended:

Stage 1. Graphically explore the emission data.

Stage 2. Assess the distributional properties of the emission data. Results from Stage 1 may guide the analysis in this stage.

Stage 3. Implement the classification algorithm, based on the conclusions from Stage 2.

Stage 4. Evaluate the ST/FTP association in terms of results obtained from Stage 3.

Stage 5. Test various hypotheses, and repeat the whole process for further adjustment.

Figure 2.4 shows a summary flow chart of the block-stages involved in the analysis.

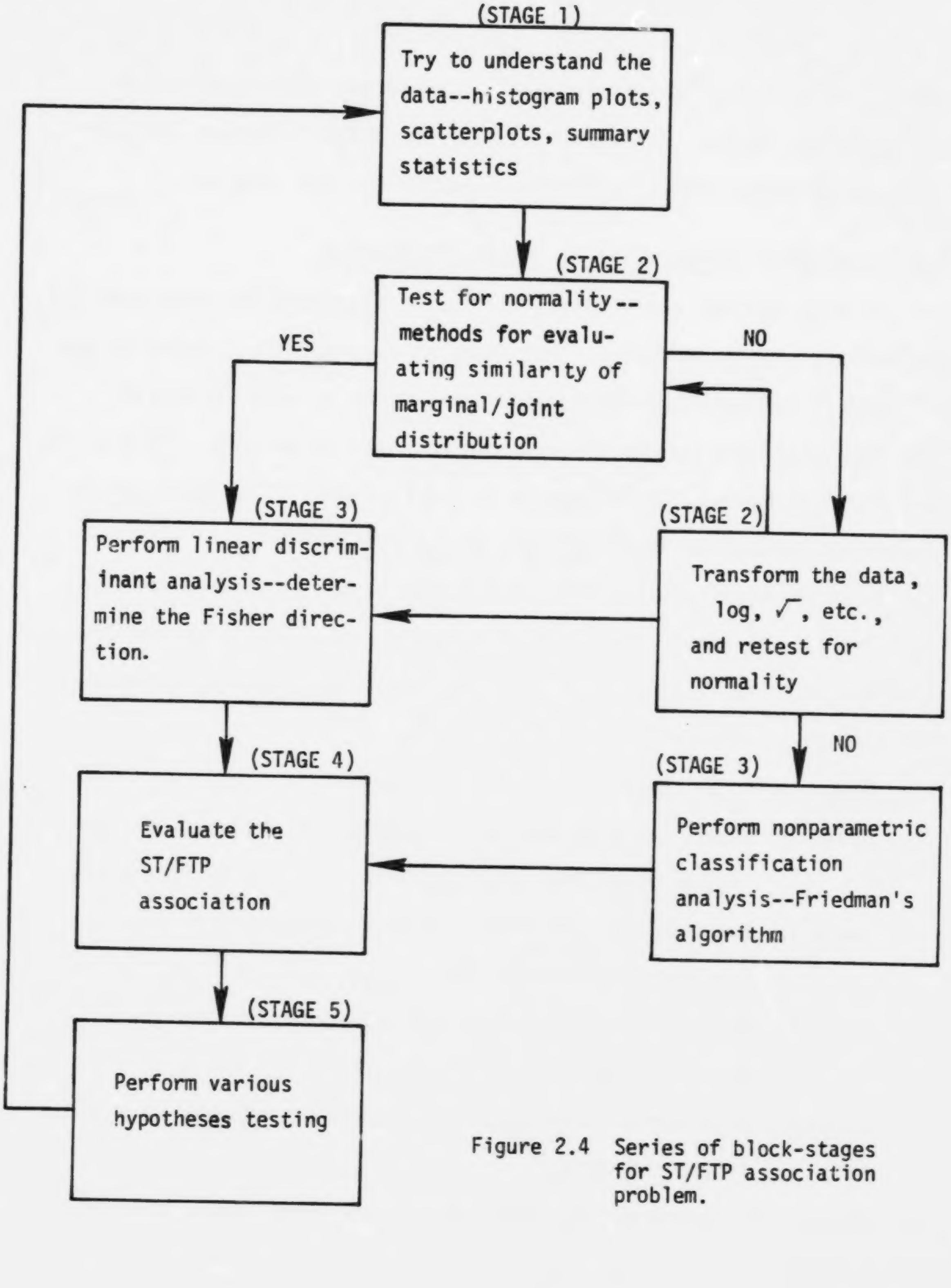

(STAGE 1)
Try to understand the data--histogram plots, scatterplots, summary statistics
(STAGE 2)
Test for normality--methods for evaluating similarity of marginal/joint distribution
YES
NO
(STAGE 3)
Perform linear discriminant analysis--determine the Fisher direction.
(STAGE 2)
Transform the data, log, √, etc., and retest for normality
NO
(STAGE 4)
Evaluate the ST/FTP association
(STAGE 3)
Perform nonparametric classification analysis--Friedman's algorithm
(STAGE 5)
Perform various hypotheses testing

Figure 2.4 Series of block-stages for ST/FTP association problem.

2.1.1 Stage 1. Graphical Exploration

Data analysis starts with an attempt to "feel what the data are like." It is instinctive for a statistician to ask the following questions:

- What does the distribution for a particular variable look like? Does it look like a normal or a lognormal distribution?
- Are there any indications of a possible relationship between a specified pair of variables?
- What is the data quality? Are there many missing values, errors, or outliers? ("Outliers" are observations which are different enough in magnitude from the rest of the observations that they are regarded as having come from another population.)

The answers to these questions can best be obtained through various plotting schemes. To be able to "see" things is the first step towards understanding. Three commonly used plotting routines are available in the computer statistical packaged programs. These routines are:

- Histogram plotting routine
- Bivariate plotting routine
- Descriptive routine regression.

Figure 2.5 shows some examples of histogram plots. From these plots, the statistician may spot some salient features of distribution and thus gain insights for a thorough analysis in the next stage.

Using the bivariate plotting routine, scatterplots, together with regression equations and their residual plots, can be printed for various pairs of emission variables, e.g., scatterplots for short-test NO_x versus short-test CO, or short-test NO_x versus FTP-NO_x. The scatterplots can

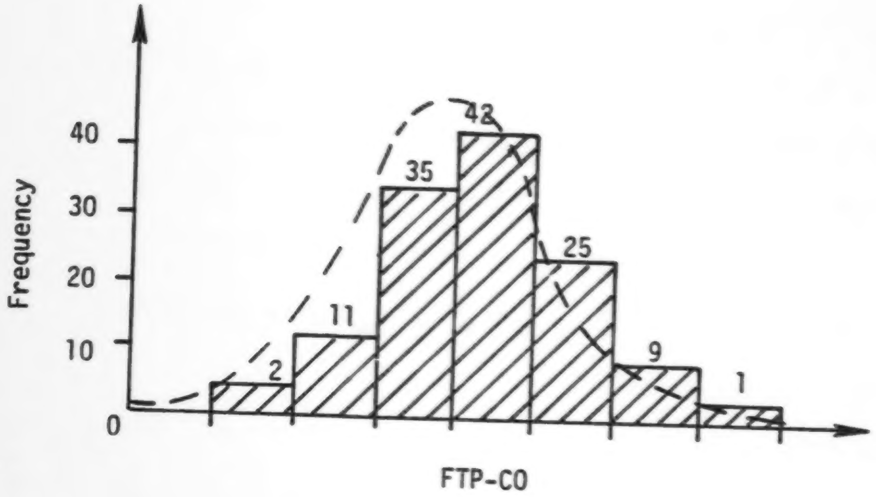


Figure 2.5a Histogram plot for FTP-CO. Suggests the distribution is normal.

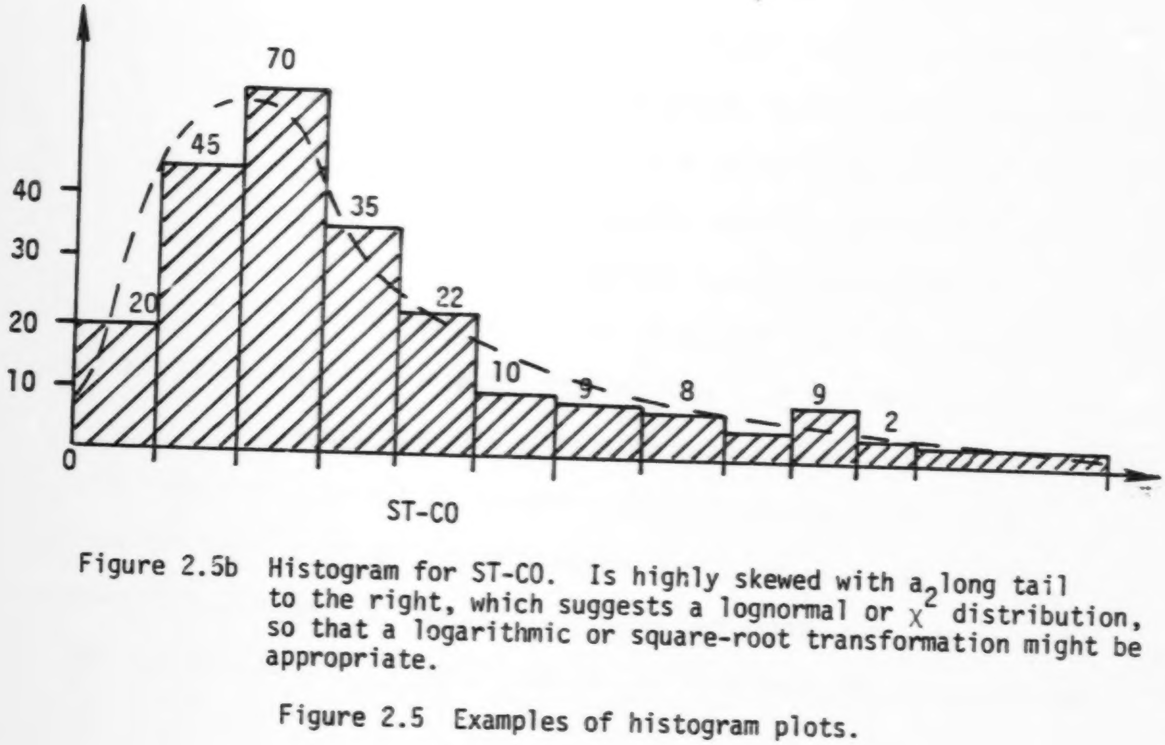
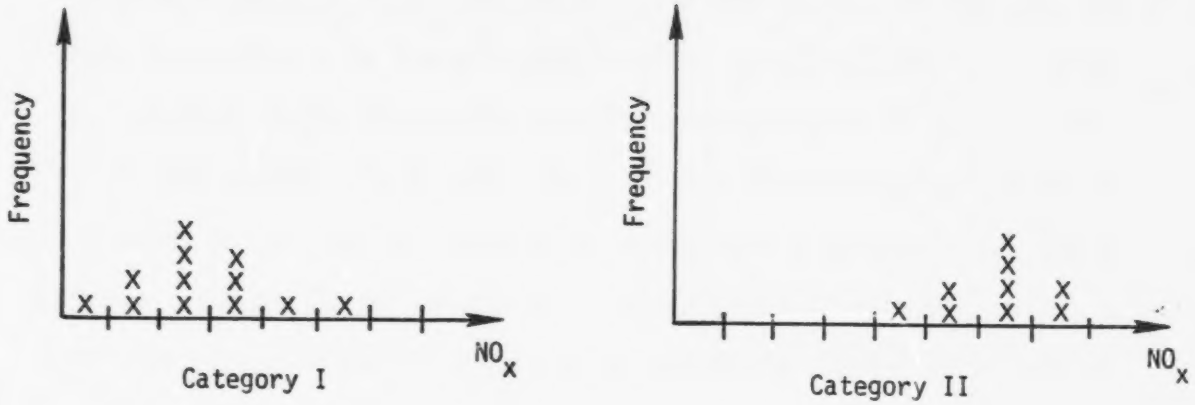


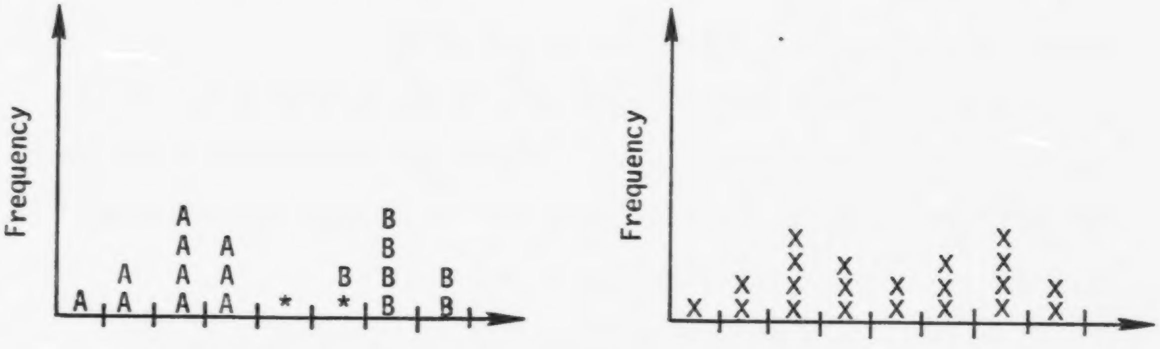
Figure 2.5b Histogram for ST-CO. Is highly skewed with a long tail to the right, which suggests a lognormal or χ^2 distribution, so that a logarithmic or square-root transformation might be appropriate.

Figure 2.5 Examples of histogram plots.



Category I

Category II



Categories I and II

Combined group

Figure 2.5c Further examples of histogram plots.

reveal any possible functional relationship between the two variables; the residual plots (obtained from fitting the linear equation) can detect outliers or strange points. Also requested from such a routine are plots for a pair of ST emission variables with points from either Category I or II, identified by distinct letters. (See Figure 2.6). Unless some of these plots indicate a good separation between the two sets of letters, there is likely to be a poor association between the ST and FTP, regardless of the various formal procedures to be pursued in the later course of the analysis.

A descriptive routine computes summary statistics--mean, covariance, largest and smallest values--and lists all data and all cases containing missing values. The quality of the data can be appraised by scanning the output, particularly for missing values and errors.

Special attention should also be paid to the covariance matrix for the six pollutant variables--namely, three ST emission variables and three FTP emission variables; small values in the off-diagonal terms indicate that the variables are mutually uncorrelated.

2.1.2 Stage 2. Assessment of the Distributional Properties of the Emission Data

An assumption of multivariate normality underlies much of the standard, classical multivariate statistical methodology. But in real data, possibility of departure from normality is prevalent. Thus, it is essential to check the normality assumption before applying the classical theory. Such a check will be helpful not only in guiding the subsequent analysis, but also in suggesting the need for a transformation

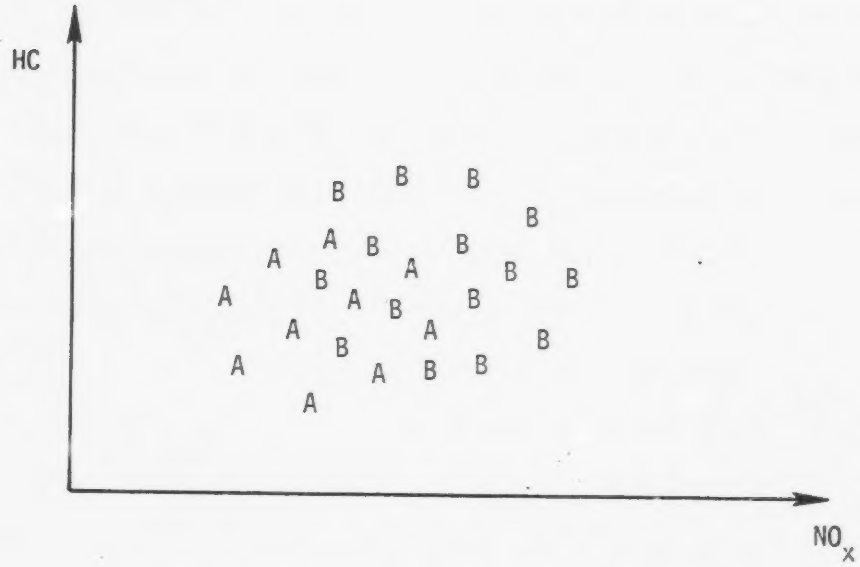


Figure 2.6 Scatterplot, with points from Categories I and II identified by distinct letters.

of the data to make them more nearly normally distributed (and therefore capable of having the classical theory applied).

The joint and marginal distributions of the emissions measured by the short test for in-use vehicles have not been checked; such a check will be necessary once the Portland data are made available.

The following steps for assessing joint normality are recommended:

Step 1. Test for marginal normality (FTP and ST separately).

Step 2. If the hypothesis of normality is rejected in Step 1, perform a transformation of the data and redo Step 1. (NOTE: If it is obvious that the data cannot be normalized, go to Stage 3.)

Step 3. Test for joint normality.

Step 1. Test for Marginal Normality. Although marginal normality does not imply joint normality, the presence of many types of non-normality is often reflected in the marginal distributions as well. Hence, a natural, simple, and preliminary step in evaluating the normality of multivariate data is to study the reasonableness of marginal normality for the observations on each of the variables.

Two commonly used techniques for assessing marginal normalities are:

- normal probability plot
- skewness and kurtosis coefficients and the Kolmogorov-Smirnov (K-S) test.

A normal probability plot is a graphical tool for assessing univariate normality. The technique consists of

1. Ordering the n observations from smallest to largest, where $X_{(i)}$ is the i^{th} smallest observation in the sample, $i=1,\ldots,n$, and $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$.
2. Letting $P_i=(i - 1/2)/n$.
3. Plotting $X_{(i)}$ against P_i on normal probability paper.

A linear configuration on such a plot would correspond to adequate normality of the observations (Figure 2.7a), while systematic and subtle departure from normality would be indicated by deviations from linearity (Figure 2.7b).

Computing Support: TSC computer library, BMDP5D (the histogram program mentioned in Stage 1).

Skewness and kurtosis coefficients and the K-S test provide a statistics-based formal test on univariate normality. Let $X_1,\ldots, X_n$ be n univariate emission observations. The skewness and kurtosis coefficients are defined as

$$\sqrt{b_1} = \frac{\sqrt{n}\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{3/2}}$$

$$b_2 = \frac{n\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{2}} ,$$

where $\overline{X} = \sum_{i=1}^{n} X_i/n$.

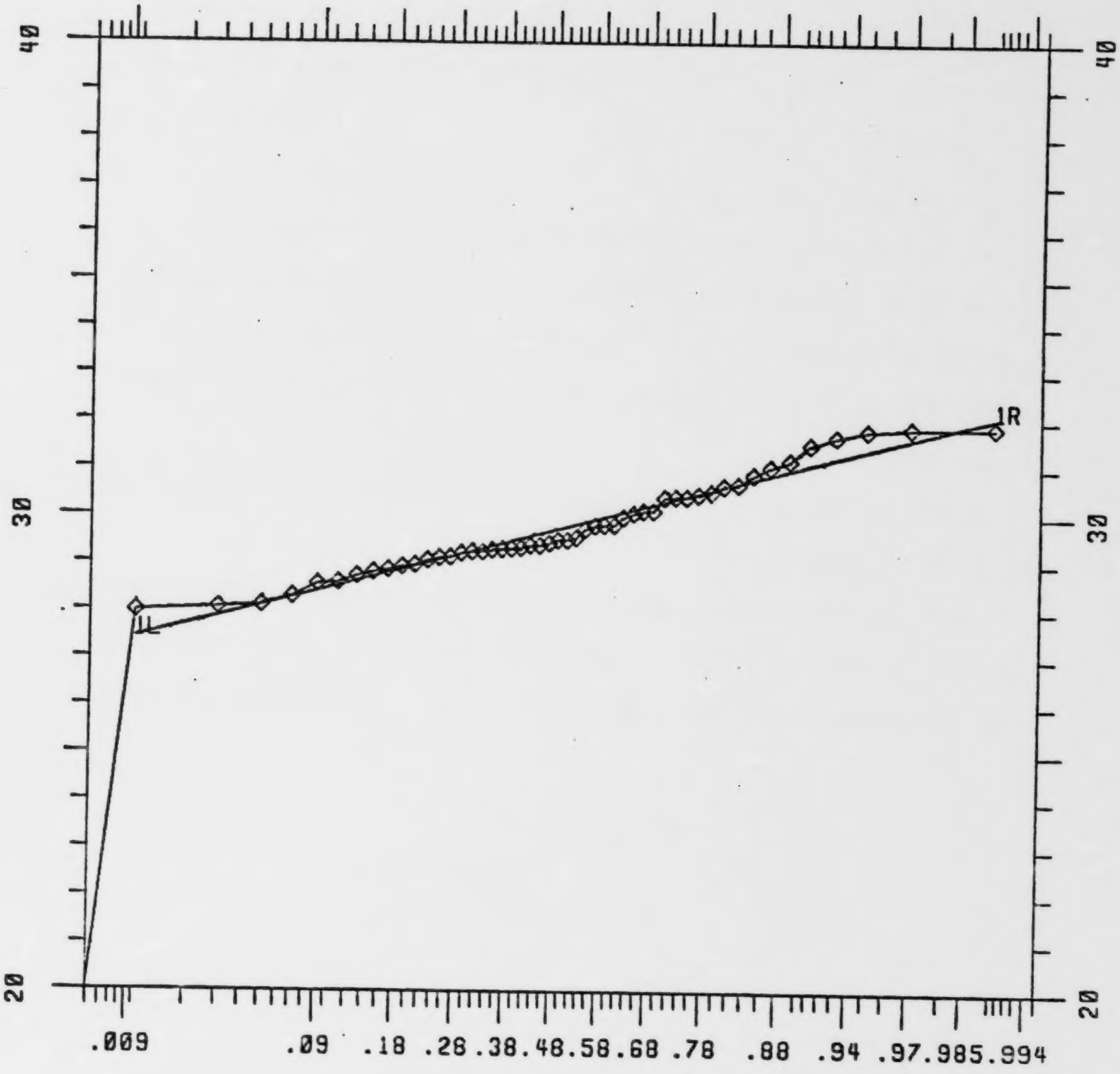


Figure 2.7a Simulated normal data plotted on normal paper.

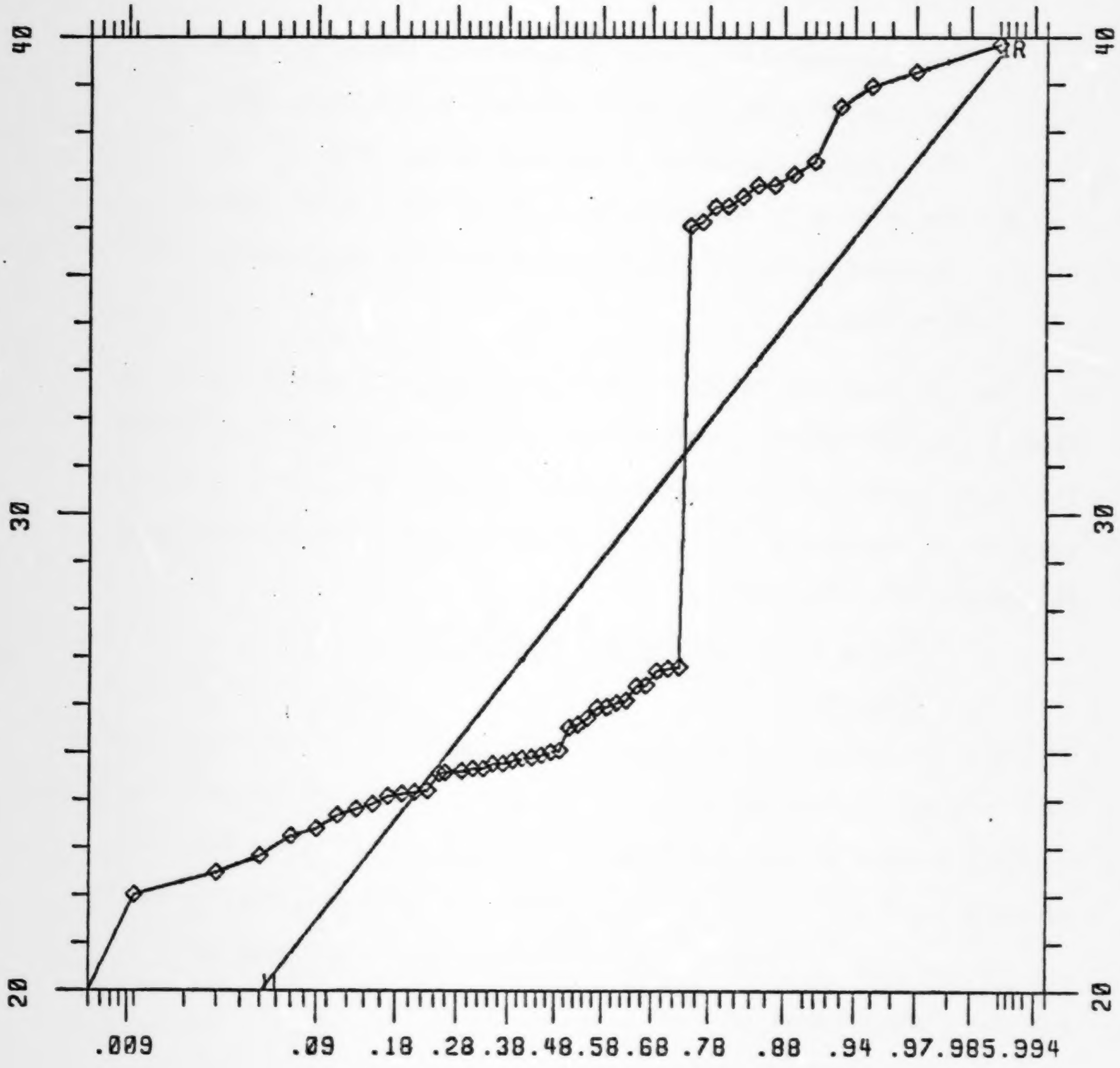


Figure 2.7b Simulated non-normal data plotted on normal paper.

The following are characteristics of $\sqrt{b_1}$ and b_2:

1. When the population density is symmetric, then $b_1=0$. When the population density has a long tail to the right, then $\sqrt{b_1} > 0$; and when it has a long tail to the left, $\sqrt{b_1} < 0$.
2. The value of b_2 for any normal distribution is $b_2=3$. But if the distribution is concentrated toward the mean, then $b_2 < 3$; otherwise, $b_2 > 3$.

For real data, the values of these coefficients are subject to random fluctuations; therefore, a statistical table has to be used to determine the value. Tables of approximately 5% and 1% points of these two statistics are given in Appendix A (Table A-1). The procedures for testing skewness and kurtosis are given below.

SKEWNESS & KURTOSIS TEST ($\sqrt{b_1}=0$ and $b_2=3$)

PROCEDURE	EXAMPLE
(1) Select a variable to be tested, say, variable X. (This may be the original variable or the transformed variable.)	(1) Suppose, from a histogram plot in Stage 1, that the CO measured from ST for all the cars passing the FTP looks normal. One then decides to test the ST-CO variable for normality.
(2) Set a significance level, α. Following the convention, set α equal to either 0.01 or 0.05.	(2) Set $\alpha=0.05$

(3) Compute the skewness coefficient, $\sqrt{b_1}$, and kurtosis coefficient, b_2, for the variable X. These two numbers can be easily obtained from the output of a frequency count subroutine such as BMDP2D or SPSS-FREQUENCY SUBPROGRAM.	(3) Using BMDP2D subroutine, one obtains $\sqrt{b_1}=0.2$ and $b_2=3.5$.
(4) Let N be the number of sample points. Look for the one-tailed percentage point, p_1, of the distribution of $\sqrt{b_1}$ determined by N and α from the table for testing skewness (see Table A-1a). Also, look for percentage point p_2 of the distribution of b_2 determined by N and α from the table for testing kurtosis (see Table A-1b).	(4) Suppose there are 112 sample points. From the table for testing skewness, one obtains the one-tailed percentage point of the distribution of $\sqrt{b_1}$ as $p_1=0.350$ (determined by 112 and 0.05). Similarly, from the table for testing kurtosis, one obtains the percentage point of the distribution of b_2 or $p_2=3.71$.
(5) If $\sqrt{b_1}$ obtained from (3) is smaller than p_1, accept the hypothesis that $\sqrt{b_1}=0$ at α significance level and thus passes the skewness test for normality. Otherwise, reject the hypothesis that $\sqrt{b_1}=0$;	(5) Since $\sqrt{b_1}=0.2 < 0.35$, it follows that the hypothesis $\sqrt{b_1}=0$ is accepted at 0.05 significance level. Also from $b_2=3.5 < 3.71$, it follows that the hypothesis $b_2=3$ is accepted at 0.05 significance level. Therefore,

X, therefore is not normal. Similarly, if b_2 obtained from (3) is smaller than p_2, accept the hypothesis that $b_2=3$ at α significance level and thus passes the kurtosis test for normality. Otherwise, reject the hypothesis that $b_2=3$; X, therefore, is not normal.

X passes the first test of normality and thus can proceed for a further check for normality as shown in the next procedure.

In the event that $\sqrt{b_1}=0$ and $b_2=3$ have been statistically accepted, which is an indication of normality, the K-S test can be utilized to further test the hypothesis that the observations are normally distributed. The procedure for the K-S test can be described as follows:

KOLMOGOROV-SMIRNOV TEST

PROCEDURE	EXAMPLE
(1) Select the variable to be tested, say, X. (This may be the original variable or the transformed variable).	(1) Suppose, in Stage 1, that the histogram of CO measured from ST for cars passing the FTP looks like normal. Then one decides to formally test the normality of the ST-CO variable for cars passing the FTP.
(2) Set a significance level α. Following the convention, set $\alpha=0.01$ or 0.05.	(2) Set $\alpha=0.05$

(3) Calculate the sample mean, $\overline{X}$, and sample variance S^2 of variable X. These two numbers can be easily obtained from the output of a descriptive subroutine such as BMDP5D or SPSS-DESCRIPTIVE SUBPROGRAM.	(3) Using SPSS-DESCRIPTIVE SUBPROGRAM, one obtains the sample mean $\overline{X}$, and sample variance, S^2, of FTP-CO variable as $\overline{X}=34.5$ and $S^2=2.0$.
(4) Set the null hypothesis, H_0: X follows a normal distribution with mean $\overline{X}$ and variance S^2, which can be written $(N(\overline{X},S^2))$. Implement the subroutine KOLMO of the SSP-1 (available through MIDAS and the TSC computing library) on X data with hypothesized distribution as stated in H_0.	(4) Set the null hypothesis as H_0: ST-CO follows a normal distribution with mean 34.5 and variance 2.0. Implement the subroutine KOLMO of the SSP-1 on ST-CO data with hypothesized distribution as stated in H_0.
(5) From the printout of the KOLMO program, look for the K-S distance, D. For the significance α-level, look up the critical value P for the K-S test of goodness-of-fit from Table A-2b (see Appendix A). (NOTE: If the number of sample points is less than 100, find the critical value from the table under the α-level column. If the number of sample points	(5) From the printout of the KOLMO program, look for the K-S distance D and find D=0.035. For the 0.05 significance level, look up the critical value P from Table A-2b. Since there are 112 sample points, the asymptotic formula for obtaining P is used and we get $P=0.886/\sqrt{(112)}=0.083$. Since P>D, accept the hypothesis H_0, i.e., X follows N(34.5,2.0).

is greater than 100, compute the critical value from the asymptotic formula.) If $D<P$, then accept the hypothesis H_0, i.e., X follows $N(\overline{X},S^2)$. Otherwise, reject the hypothesis, i.e., X does not follow the normal distribution.

Step 2. Transformation of the Data. If, for a particular emission variable, the hypothesis of normality is rejected at Step 1, then a transformation of the variable that induces normality should be considered. Logarithmic transformations and square-root transformations are the two commonly used types.

EXAMPLE: The histogram plot of the HC values measured from the short test is highly skewed with a long tail to the right (Figure 2.5b, p. 2-12); thus, the logarithms of individual HC observations are utilized, and the resultant transformed sample histogram now looks normal. Going again through Step 1, the hypothesis of normality is accepted. Therefore, the HC variable follows a lognormal distribution. Alternatively, the square-root transformation can be used, and the acceptance of the normality of the transformed variable will mean the original variable follows a chi-square distribution.

Step 3. Test for Joint Normality. After marginal normality is confirmed for all the individual emission variables (or transformed variables), the test for joint normality can be implemented for a final check.

Since there are only three emission variables, it usually suffices to test the bivariate normality for each pair of variables.

Two methods for assessing the joint normality are:

- A test based on linear regressions
- A test based on radius-and-angle representation of multivariate data.

A property of the multivariate normal distribution is that the regression of each variable on all others is linear. Thus, a simple test for detecting joint non-normality would be to perform a linear regression line fit and decide from the residual plot whether the regression shows any sign of curvature. Figure 2.8 illustrates two possible situations. If the residual plot indicates obvious curvature, reject the hypothesis of joint normality; otherwise,accept bivariate normality.

Computing Support: Any linear regression program with a residual plot will do.

The second procedure is based on two things: 1) representing a two-dimensional data point in terms of radius and angle; and 2) under the joint normal assumption, the radius's following a chi-squared distribution with 2 degrees of freedom and the angle's following a uniform distribution over $(-\pi/2,\pi/2)$. Specifically, the operations of this procedure can be described as follows:

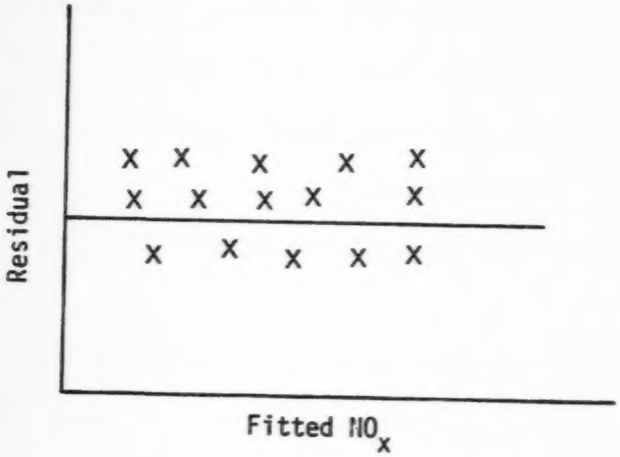


Figure 2.8a Case for the acceptance for possible joint normality of NO_x and CO.

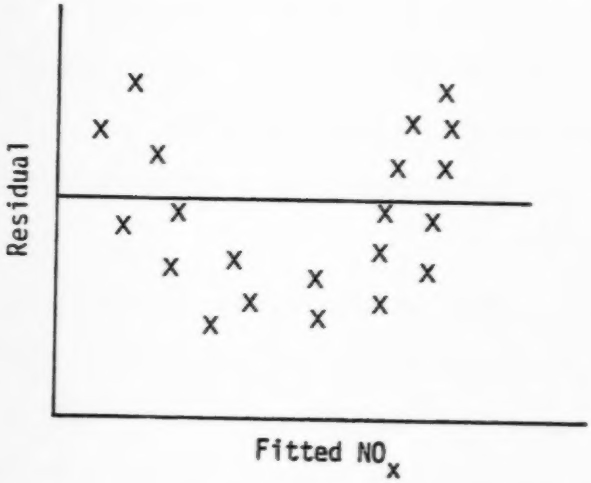


Figure 2.8b Residual plot indicating curvature; the hypothesis of joint normality is rejected.

Figure 2.8 Linear regression line fit for assessing the joint normality of NO_x and CO, where $\widehat{NO}_x = a + b(CO)$ and the residual = $NO_x - \widehat{NO}_x$.

PROCEDURE	EXAMPLE
(1) Let $\underline{X}_1, \ldots \underline{X}_n$ be n two-dimensional observations $\underline{X}_k = (X_{1k}, X_{2k})$. Compute the covariance matrix S where $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix}$ $S_{ij} = \frac{1}{n-1}\left[\left(\sum_{k=1}^{n} X_{ik} X_{jk}\right) - \left(\sum_{k=1}^{n} X_{ik}\right)\left(\sum_{k=1}^{n} X_{jk}\right)\Big/ n\right]$ and the mean vector $\underline{\overline{X}}$ $\underline{\overline{X}} = (\overline{X}_1, \overline{X}_2)$, where $\overline{X}_i = \frac{\sum_{k=1}^{n} X_{ik}}{n}$.	(1) From Table 2.1, which lists HC and NO_2 data, $n = 47$ $S_{11} = 1.94570$ $S_{22} = 8.52141$ $S_{12} = 0.6301$ $\overline{X}_1 = 2.84383$ $\overline{X}_2 = 3.76447$
(2) Find θ_k and r_k^2 for each two-dimensional observation. where $\theta_k = \tan^{-1}\left(\frac{Z_{2i}}{Z_{1i}}\right)$ (remember to account for the correct quadrant) and $r_k^2 = Z_{1k}^2 + Z_{2k}^2$, where	(2) Transform the data $Z_{1k} = \frac{(X_{1k}-2.84383)8.5214-(X_{2k}-3.76447)0.63}{11.74320}$ $Z_{2k} = \frac{(X_{2k}-3.76447)}{2.91915}$ The values appear in Table 2.1.

$$Z_{1k} = \frac{(X_{1k}-\overline{X}_1)S_{22} - (X_{2k}-\overline{X}_2)S_{12}}{\sqrt{S_{22}}\sqrt{S_{11}\ S_{22} - S_{12}^2}}$$

$$Z_{2k} = \frac{(X_{2k} - \overline{X}_2)}{\sqrt{S_{22}}}$$

(3) Compute SR and SLR

$$SR = \sum_{k=1}^{n} r_k^2 \text{ and}$$

$$SLR = \sum_{k=1}^{n} r_k^2 \log_e(r_k^2)$$

Calculate the goodness-of-fit statistic G

$$G = \left| \frac{\sqrt{n}}{0.887} \left(\frac{SLR}{SR} - \log_e(SR) + \log_e(n) - 0.4228 \right) \right| .$$

G has a standard normal distribution.

(4) Sort the θ values and transform them to the (0,1) interval. The transformation is:

$$0.5 + \frac{\theta_k}{2\pi} .$$

Compute the cumulative frequency distribution of the k^th ordered variable

$$\frac{k}{n+1} .$$

(3) Compute SR, SLR and G

SR = 243.786

SLR = 874.178

G = 1.517

(4) Table 2.2 gives the sorted $0.5 + \frac{\theta_k}{2\pi}$ values, $\frac{k}{n+1}$ and KS_k.

KS = 0.34862

Calculate the Kolmogorov-Smirnov distance for each value

$$KS_k = \left| 0.5 + \frac{\theta_k}{2\pi} - \frac{k}{n+1} \right|$$

and let KS be the maximum, i.e.,

$$KS = \max_k \left| 0.5 + \frac{\theta_k}{2\pi} - \frac{k}{n+1} \right| .$$

(5) Check G and KS to see if there is a significant difference.

(5) G = 1.517 and KS = 0.34862 from the normal table (Table A-3). One finds that at 0.05 significance level, G is insignificant, i.e., the distribution of r^2 is χ^2. However, from the K-S two-sample table (Table A-2a), one finds that at 0.05 significance level, KS is significant, i.e., the hypothesis that the distribution of θ is uniformly distributed is rejected. Therefore, the data is not bivariate normal. Transformation of the variable may be needed.

(6) Plot r(i) against $1-P_i$ on semilog probability paper, Figure 2.9a (the actual paper is presented as Table A-7) and plot $\theta_{(i)}$ against P_i on uniform probability paper, Figure 2.9b. If the data conforms statistically to the null hypothesis of bivariate normality, the configuration on these two probability plots should be reasonably linear. Departure from linearity on either or both of the plots would indicate specific types of departure from the null hypothesis.

Computing Support: TSC computer library.

TABLE 2.1 BIVARIATE HC AND NOX VALUES WITH ASSOCIATED Z_1, Z_2, θ, r^2 AND $r^2\log_e r^2$

HC	NOX	Z_1	Z_2	θ	r^2	$r^2\log_e r^2$
4.8	4.44	1.38324	0.23141	0.16576	1.96691	1.33054
2.64	7.00	-0.32152	1.10838	1.85312	1.33188	0.38171
6.35	1.84	2.64750	-0.65926	-0.24405	7.44388	14.94279
2.53	2.57	-0.16364	-0.40918	-1.95123	0.19421	-0.31827
2.02	2.20	-0.51387	-0.53593	-2.33517	0.55120	-0.32829
2.68	1.82	-0.01455	-0.66611	-1.59263	0.44391	-0.36051
15.85	1.02	9.58514	-0.94016	-0.09777	92.75883	420.19779
2.52	1.49	-0.11295	-0.77916	-1.71475	0.61984	-0.29647
1.82	1.12	-0.60105	-0.90591	-2.15658	1.18192	0.19755
11.66	4.57	6.35420	0.27595	0.04340	40.45202	149.67719
14.15	4.84	8.14657	0.36844	0.04520	66.50241	279.12646
3.45	1.69	0.55117	-0.71064	-0.91111	0.80881	-0.17163
3.31	1.57	0.45602	-0.75175	-1.02553	0.77309	-0.19897
3.30	3.01	0.37150	-0.25846	-0.60784	0.20481	-0.32476
2.45	3.64	-0.27910	-0.04264	-2.98999	0.07972	-0.20163
4.50	5.30	1.11940	0.52602	0.43929	1.52976	0.65032
3.43	1.44	0.55008	-0.79628	-0.96627	0.93665	-0.06130
2.76	2.51	0.00648	-0.42974	-1.55572	0.18472	-0.31197
1.15	2.25	-1.14786	-0.51881	-2.71709	1.58674	0.73257
3.48	3.21	0.49139	-0.18994	-0.36885	0.27754	-0.35575
3.57	6.30	0.39089	0.86859	1.14791	0.90724	-0.08832
3.46	5.55	0.35132	0.61166	1.04944	0.49755	-0.34732
1.26	2.33	-1.07233	-0.49140	-2.71189	1.39137	0.45955
3.01	2.70	0.17770	-0.36465	-1.11736	0.16455	-0.29693
1.97	3.46	-0.61775	-0.10430	-2.97433	0.39250	-0.36707
3.26	2.33	0.37896	-0.49140	-0.91387	0.38509	-0.36748
3.62	1.77	0.67024	-0.68324	-0.79500	0.91604	-0.08033
1.05	1.80	-1.19628	-0.67296	-2.62917	1.88396	1.19326
2.20	1.47	-0.34408	-0.78601	-1.98342	0.73620	-0.22546
1.63	1.61	-0.76521	-0.73805	-2.37426	1.13026	0.13840
1.61	1.75	-0.78723	-0.69009	-2.42186	1.09596	0.10043
2.45	3.55	-0.27427	-0.07347	-2.87986	0.08062	-0.20301
2.51	2.89	-0.19532	-0.29956	-2.14859	0.12789	-0.26302
4.70	3.48	1.36219	-0.09745	-0.07142	1.86505	1.16246
3.31	2.60	0.40076	-0.39891	-0.78309	0.31973	-0.36458
2.73	2.99	-0.04104	-0.26531	-1.72400	0.07207	-0.18956
3.34	2.81	0.41126	-0.32697	-0.67171	0.27604	-0.35532
6.07	2.04	2.43359	-0.59074	-0.23814	6.27133	11.51409

TABLE 2.1--Cont'd

HC	NOX	Z_1	Z_2	θ	r^2	$r^2 \log_e r^2$
3.03	2.62	0.19650	-0.39206	-1.10618	0.19232	-0.31706
4.64	3.54	1.31543	-0.07690	-0.05839	1.73627	0.95796
3.53	3.53	0.51050	-0.08032	-0.15606	0.26706	-0.35259
3.25	2.07	0.38566	-0.58047	-0.98437	0.48567	-0.35076
3.33	2.11	0.44156	-0.56677	-0.90894	0.51620	-0.34134
3.30	1.51	0.45199	-0.77230	-1.04130	0.80075	-0.17794
3.77	5.21	0.59451	0.49519	0.69450	0.59865	-0.30715
3.09	2.49	0.24702	-0.43659	-1.05591	0.25163	-0.34720
2.39	1.62	-0.21426	-0.73462	-1.85458	0.58558	-0.31338

TABLE 2.2 SORTED VALUES OF U = $0.5 + \frac{\theta_k}{2\pi}$, THE ASSOCIATED CUMULATIVE FREQUENCY F = $\frac{k}{n+1}$, AND THE KS STATISTIC

θ	U	F	KS
-2.98999	0.02413	0.02083	0.00329
-2.97433	0.02662	0.04167	0.01505
-2.87986	0.04166	0.06250	0.02084
-2.71709	0.06756	0.08333	0.01577
-2.71189	0.06839	0.10417	0.03578
-2.62917	0.08155	0.12500	0.04345
-2.42186	0.11455	0.14583	0.03128
-2.37426	0.12212	0.16667	0.04454
-2.33517	0.12835	0.18750	0.05915
-2.15658	0.15677	0.20833	0.05156
-2.14859	0.15804	0.22917	0.07113
-1.98342	0.18433	0.25000	0.06567
-1.95123	0.18945	0.27083	0.08138
-1.85458	0.20483	0.29167	0.08683
-1.72400	0.22562	0.31250	0.08688
-1.71475	0.22709	0.33333	0.10624
-1.59263	0.24653	0.35417	0.10764
-1.55572	0.25240	0.37500	0.12260
-1.11736	0.32217	0.39583	0.07367
-1.10618	0.32395	0.41667	0.09272
-1.05591	0.33195	0.43750	0.10555
-1.04130	0.33427	0.45833	0.12406
-1.02553	0.33678	0.47917	0.14238
-0.98437	0.34333	0.50000	0.15667
-0.96627	0.34621	0.52083	0.17462
-0.91387	0.35455	0.54167	0.18711
-0.91111	0.35499	0.56250	0.20751
-0.90894	0.35534	0.58333	0.22800
-0.79500	0.37347	0.60417	0.23069
-0.78309	0.37537	0.62500	0.24963
-0.67171	0.39309	0.64583	0.25274
-0.60784	0.40326	0.66667	0.26341
-0.36885	0.44130	0.68750	0.24620
-0.24405	0.46116	0.70833	0.24718
-0.23814	0.46210	0.72917	0.26707
-0.15606	0.47516	0.75000	0.27484
-0.09777	0.48444	0.77083	0.28639

TABLE 2.2--Cont'd

θ	U	F	KS
-0.07142	0.48863	0.79167	0.30303
-0.05839	0.49071	0.81250	0.32179
0.04340	0.50691	0.83333	0.32643
0.04520	0.50719	0.85417	0.34697
0.16576	0.52638	0.87500	0.34862
0.43929	0.56992	0.89583	0.32592
0.69450	0.61053	0.91667	0.30613
1.04944	0.66702	0.93750	0.27048
1.14791	0.68270	0.95833	0.27564
1.85312	0.79493	0.97917	0.18423

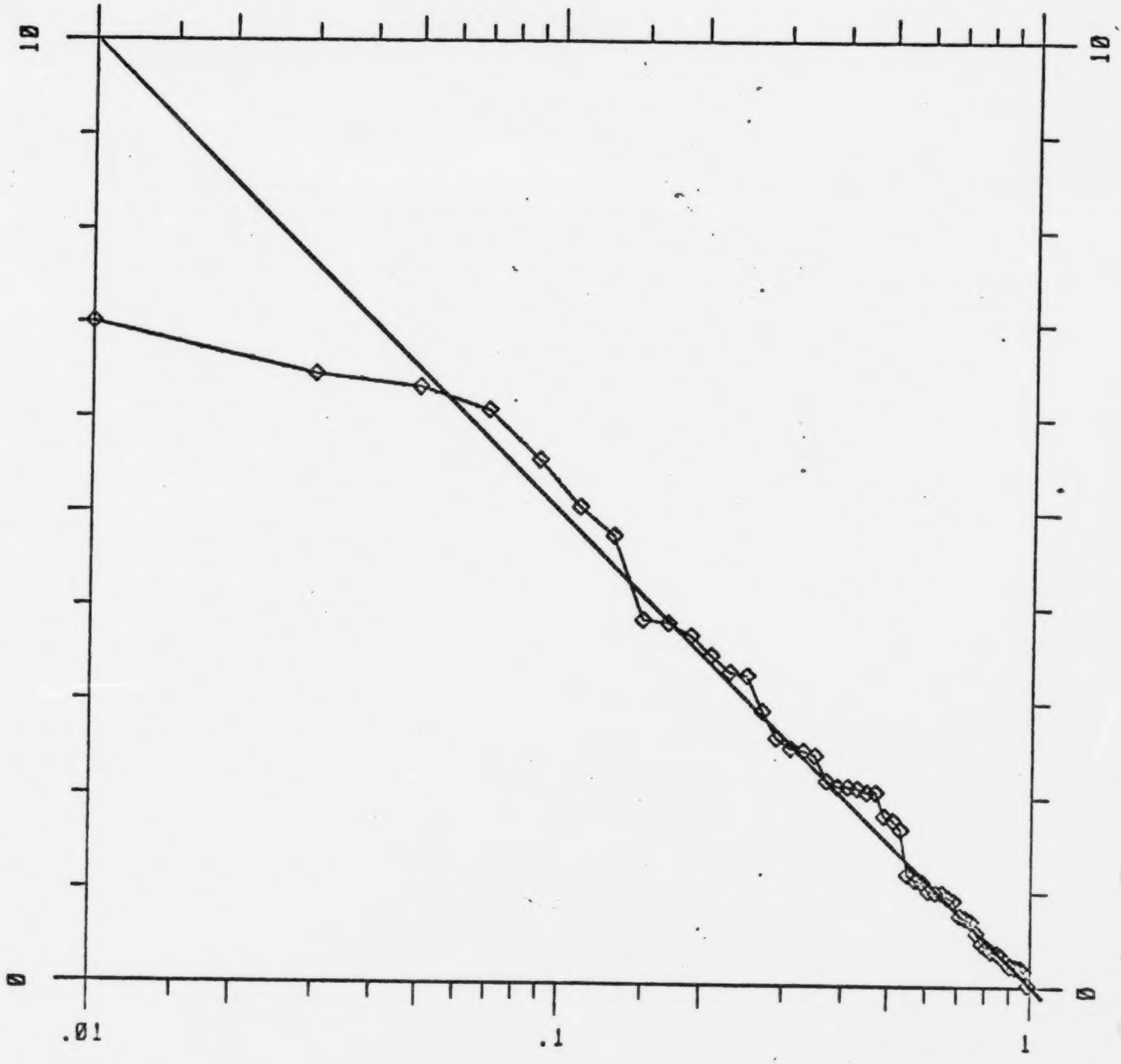


Figure 2.9a Distance representation of bivariate normal data on a semilog probability plot.

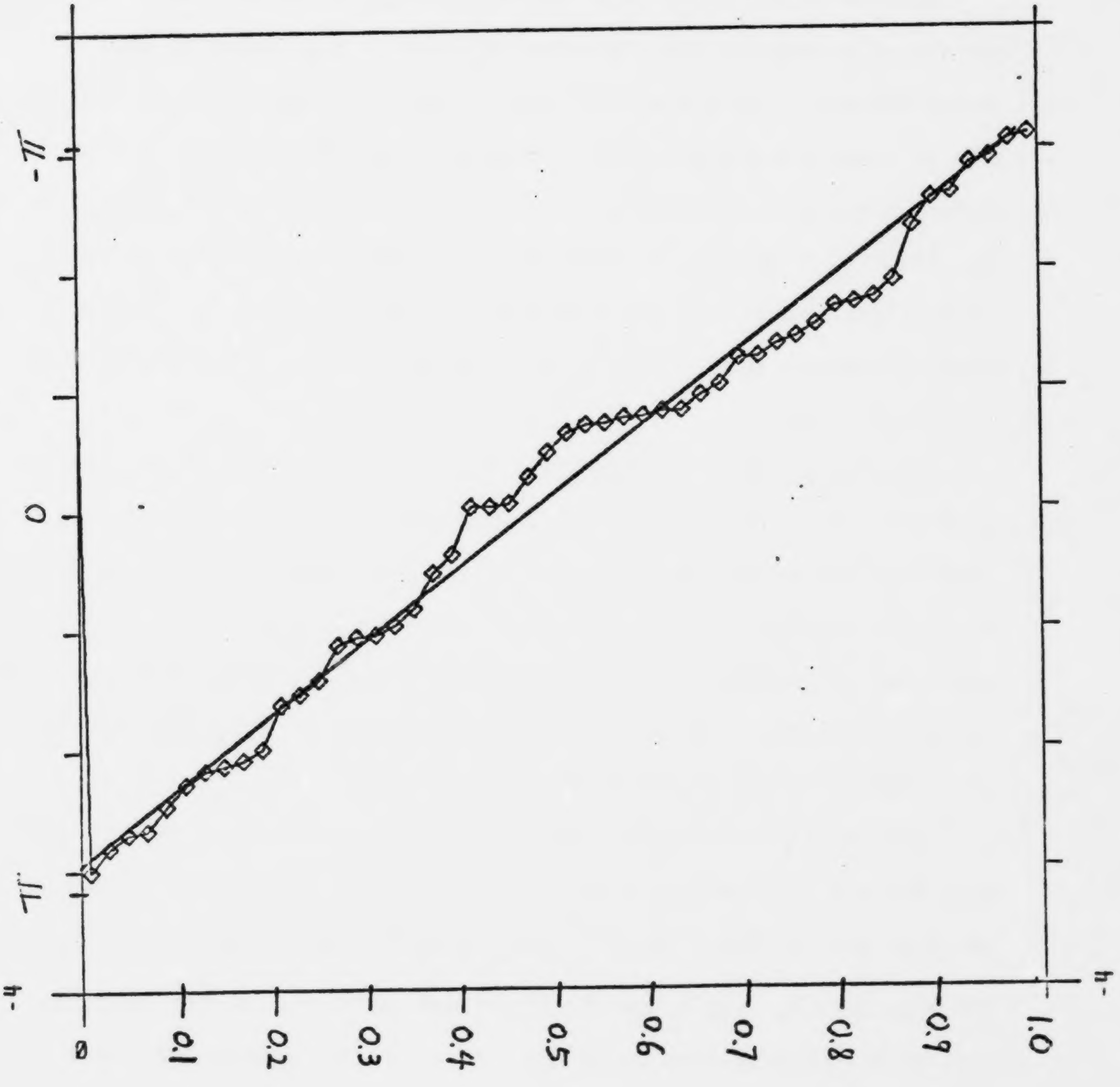


Figure 2.9b Angle representation of bivariate normal data on a uniform probability plot.

2.1.3 Stage 3. Implementation of Classification Algorithms

Analysis of the ST/FTP association involves the use of classification techniques to separate the short-test emission data according to known FTP categorizations. Since classical classification schemes are highly sensitive to underlying distributions, the conclusion from Stage 2 is a crucial factor in deciding which particular classification technique to adopt. Specifically, if Stage 2 concludes that the underlying distributions are multivariate normal, then the classical discriminant analysis algorithm is used. Otherwise, the Friedman's recursive partitioning nonparametric classification is used.

Discriminant Analysis Algorithm. This algorithm produces a linear rank score, the Fisher discriminant function, which separates the emission data space into two parts, "pass" and "fail". If, as concluded from Stage 2, each of the two batches of points follows a multivariate normal distribution, then the Fisher discriminant function computed from the algorithm is optimal; that is, no other linear rank procedure will generate a lower expected probability of misclassification [Anderson 1958, pp. 129-136].

Specifically, assume that the observations of short-test emission vectors for cars in Categories I and II are $Y_1,\ldots,Y_{N_1}$ from a multivariate normal distribution $N(M_Y,S)$ and $Z_1,\ldots,Z_{N_2}$ from $N(M_Z,S)$, respectively, where the true means M_Y and M_Z and the covariance matrix are unknown (and thus have to be estimated from the data). The algorithm computes the Fisher discriminant function according to the following steps:

1. Compute the sample mean vectors, $\overline{Y}$ and $\overline{Z}$, as well as the sample covariance matrices, S_Y and S_Z.

2. Compute the pooled covariance matrix as

$$S = [(N_1-1)S_Y + (N_2-1)S_Z]/(N_1+N_2-2) \quad .$$

3. Form the Fisher discriminant function (or the Fisher linear rank score) as

$$r(x) = xS^{-1}(\overline{Y}-\overline{Z})^T - 1/2(\overline{Y}+\overline{Z})S^{-1}(\overline{Y}+\overline{Z})^T \quad ,$$

where $x = (X_1, X_2, X_3)$ is an emission testing vector. NOTE: $r(x)$ is a linear combination of the three individual emission variables, X_1, X_2, and X_3. T designates the transpose of the matrix.

4. The desired cutoff point is given by the number

$$C^* = \log_e(N_2/N_1) \quad .$$

Thus, the best way to classify the car with emission vector x is to assign it to Category I if $r(x) \geq C^*$ and to Category II if $r(x) < C^*$. NOTE: If one multiplies both sides of the inequality by -1, the inequality will reverse direction.

Since the Fisher discriminant function produces two kinds of classification errors, inclusion of "good" vehicles in the "fail" group and of "bad" in the "pass" group is possible. Any desired subjective weighting of the relative undesirabilities of these two kinds of errors can be incorporated into the discrimination of the cutoff point. For example, let C_1 be the cost of classifying a car from Category I into Category II and C_2 be the cost of classifying a Category II car into Category I. The cutoff point for the Fisher discriminant function, then, should be given by

$$C^* = \log_e(N_2C_2/N_1C_1) \quad .$$

EXAMPLE: Assume that the raw data contain $N_1 = 70$ cars from Category I and $N_2 = 43$ cars from Category II. Suppose that Stage 2 confirmed that, for each category, the short-test emission variables $X_1 = \log(NO_x)$, $X_2 = CO$, and $X_3 = HC$ have a multivariate normal distribution. Then the Fisher discriminant function can be computed, using linear discriminant analysis, according to the following computational steps:

1. Compute the sample mean and sample covariance of Category I and of Category II. Assume, for this example, that these values are

$$\overline{\underset{\sim}{Y}} = (1, 8, 2), \text{ i.e., } \overline{X}_1 = 1,\ \overline{X}_2 = 8,\ \overline{X}_3 = 2$$

$$S_Y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\overline{\underset{\sim}{Z}} = (2,10,1)$$

$$S_Z = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad .$$

2. Compute the pooled covariance matrix. In this example,

$$S = [(70-1)S_Y + (43-1)S_Z]/(70+43-2)$$

$$= \frac{1}{111} \cdot \left(\begin{bmatrix} 69 & 0 & 0 \\ 0 & 138 & 0 \\ 0 & 0 & 207 \end{bmatrix} + \begin{bmatrix} 126 & 0 & 0 \\ 0 & 84 & 0 \\ 0 & 0 & 42 \end{bmatrix} \right) = \begin{bmatrix} 1.76 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2.24 \end{bmatrix}$$

3. Compute the matrix inverse of S, S^{-1}. Here,

$$S^{-1} = \begin{bmatrix} .57 & 0 & 0 \\ 0 & .50 & 0 \\ 0 & 0 & .45 \end{bmatrix} .$$

4. Compute the Fisher discriminant function (the Fisher linear rank score). In this example, it is

$$r(x) = [X_1 X_2 X_3]\ S^{-1} \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix} - \frac{1}{2} [3\ 18\ 3]\ S^{-1} \begin{bmatrix} 3 \\ 18 \\ 3 \end{bmatrix}$$

$$= [.57X_1\ \ .50X_2\ \ .45X_3] \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix} - [0.86\ 4.5\ .67] \begin{bmatrix} 3 \\ 18 \\ 3 \end{bmatrix}$$

$$= -.57X_1 - 1.0X_2 + .45X_3 - 85.59 ,$$

where $x = (X_1, X_2, X_3)$ is a testing emission vector.

5. Find the cutoff point for the Fisher linear rank score. Here, it is

$$C^* = \log(43/70) = -0.49 .$$

Thus, a car with short-test emission vector x is classified into Category I (passing the FTP) if $r(x) \geq -0.49$, or equivalently,

$$-0.57[\log(NO_x)] - 1.0(CO) + 0.45(HC) \geq 85.1 ,$$

and to Category II if $r(x) < -0.49$.

If a false negative error were twice as crucial as a false positive error, the derived cutoff point for the Fisher linear rank score would be

$$C^* = \log_e(2 \times 43/70) = .21 \quad .$$

In practice, some of the variables least crucial to the classification can be deleted totally from the discriminant function. When such deletion occurs, a stepwise discriminant analysis algorithm can be used to identify those variables which "best" discriminate between the two categories and to produce the Fisher discriminant function based on those variables.

Most computer statistical packages (e.g., BMDP, SPSS) contain a stepwise discriminant analysis procedure similar to that outlined above. The following example illustrates the use of such a procedure. By following the instructions listed in the example, a user with a BMDP or SPSS manual should be able to perform the analysis and interpret the results.

EXAMPLE: Assume that the short-test emission variables $X_1 = \log_e(NO_x)$, $X_2 = CO$, and $X_3 = HC$ have a multivariate normal distribution and that the raw data contain $N_1 = 70$ cars from Category I and $N_2 = 43$ cars from Category II. The subroutine BMDP7M or SPSS-DISCRIMINANT can be used to perform the stepwise discriminant analysis.

For both subroutines, a description of the card preparation, details of the printout, and interpretation of the printout are given below.

SPECIAL PROGRAM CARD PREPARATION

BMDP7M -- The program card deck consists of a number of paragraphs including PROBLEM, INPUT, VARIABLE, and GROUP. Within each paragraph is a set of (required or optional) statements to be specified by the user. The details of the card preparation are given in the BMDP manual. Statements which require special instructions are listed below.

STATEMENT	EXAMPLE
(1) In the GROUP paragraph, set PRIOR = $N_1/(N_1+N_2)$, $N_2/(N_1+N_2)$.	(1) Compute 70/(70+43) = 0.62 43/(70+43) = 0.38 .
(1)' In case costs C_1 and C_2 are associated with the false negative error and false positive error, respectively, set PRIOR = $C_1N_1/(N_1+N_2)$, $C_2N_2/(N_1+N_2)$.	(1)' If the false negative error is twice as crucial as the false positive error, set PRIOR = 0.62, 0.38.
(2) In the DISCRIMINANT paragraph, ENTER = limiting value of F-to-enter REMOVE = limiting value of F-to-remove.	(2) In the DISCRIMINANT paragraph, ENTER = 0.01 REMOVE = 0.005.
(3) In the DISCRIMINANT paragraph, FORCE = a number indicating the level through which variables are forced into the equation; a zero indicates no forcing. Assumed value is zero.	(3) For this example, no forcing is specified. But if the user decides that, e.g., variable X_2 must be included in the equation, then write FORCE = 0, 1, 0

SPSS-DISCRIMINANT--The program cards consist of a number of specification phrases followed by a number of option numbers. Details of the program are given in the SPSS manual. Statements which require special instructions are listed below.

STATEMENT	EXAMPLE
(1) ANALYSIS = lists dicriminating variables to be used in the analysis. Inclusion level may be specified. Default inclusion level is 1.	(1) ANALYSIS = X_1, X_2, X_3/ In the case where X_2 is forced to be included in the equation, write ANALYSIS = X_1,X_2(1), X_3/
(2) METHOD = $\left\{\begin{array}{l}\text{DIRECT}\\ \text{WILKS}\end{array}\right.$	(2) Write METHOD = WILKS/ If the Fisher discriminant function is to be computed, forcing all three variables in the equation, set METHOD = DIRECT/
(3) FIN = limiting value of F-to-enter FOUT = limiting value of F-to-remove	(3) FIN = 0.01 / FOUT = 0.005 /
(4) OPTION = 12 gives print classification function coefficients.	(4) OPTION = 12 / is required.

The main parts of the printouts are displayed in the OUTPUT EXAMPLE (BMDP7M) and OUTPUT EXAMPLE (SPSS-DISCRIMINANT).

PRINTOUTS

OUTPUT EXAMPLE (BMDP7M)

```
   Step number 1
                  .
                  .
                  .

   F-matrix   degree of freedom = 1     111

   Classification functions

               Group = 1        2

   Variable

     X1            0.262      0.141

   Constant     -11.627     -3.383

   Step number 2
                  .
                  .
                  .

   F-matrix   degree of freedom = 2      110

   Classification functions

               Group = 1        2

   Variable

     X1            0.690      0.544

     X3          211.621    199.016

   Constant    -213.956   -182.326
```

```
      Step number 3
                   .
                   .
                   .
      F-matrix    degree of freedom = 3       109
      Classification functions
         Group = 1        2
      Variable
        X1        0.686      0.535

        X3        0.126      0.324

        X2      211.481    198.654

 Constant      -214.141   -183.557

      Classification matrix

                                                  Number of Cases
          Group      Percentage Correct       Classified into Group

                                                  1             2
            1              95.7                  67             3
            2              83.7                   7            36
          TOTAL            89.7                  74            39
```

Summary table

Step Number	Variable Entered	Variable Removed	F-Value to Enter or Remove	U-Statistics	Approximate F-Statistics
1	X_1		131.41	0.458	131.41
2	X_3		17.35	0.396	84.06
3	X_2		2.55	0.340	63.58

OUTPUT EXAMPLE (SPSS-DISCRIMINANT)

```
Step 1
         .
         .
         .

F-matrix - degrees of freedom:    1    111

Classification function coefficients

                    Group 1        Group 2

     X1               0.262          0.141

     Constant       -11.627         -3.383

Step 2
        .
        .
        .

F-matrix - degrees of freedom:   2    110

Classification function coefficients

                    Group 1        Group 2

     X1               0.690          0.544

     X3             211.621        199.016

     Constant      -213.956       -182.326

Step 3
         .
         .
         .

F-matrix - degrees of freedom:   3    109

Classification function coefficients

                    Group 1        Group 2

     X1               0.686          0.535

     X3               0.126          0.324

     X2             211.481        198.654

     Constant      -214.141       -183.557
```

Prediction Results:

Actual Group	No. of Cases	Predicted Group Membership Gp. 1	 Gp. 2
Group 1	70	67	3
		95.7%	4.3%
Group 2	43	7	36
		12.3%	83.7%

Summary Table:

Step Number	Variable Entered	 Removed	F to Enter or Remove	Wilks' Lambda	Rao's V
1	X_1		131.41	0.458	131.41
2	X_3		17.35	0.396	84.06
3	X_4		9.55	0.340	63.58

INTERPRETATION OF COMPUTER PRINTOUT

Procedure for choosing the best subset and the accompanying Fisher discriminant function.

PROCEDURE	EXAMPLE
(1) Choose α, the significance level of choosing the best subset of the variables.	(1) Set $\alpha = .05$ $1-\alpha = .95$
(2) Look up the F-matrix of each step for the degrees of freedom d_1, d_2, and copy them.	(2) The degrees of freedom appearing in the F-matrix following each step number are recorded as follows:
(3) Look up $F_{1-\alpha}$ for 1 and d_2-1 degrees of freedom in F-table (see Table A-5), for each step number. Here, d_2 degrees of freedom are found in the table given in (2).	(3) Look up $F_{.95}(1, d_2-1)$ in the F-table for each step number and record them, Here, d_2 degrees of freedom are found in the table given in (2).

Example (2):

Step number	Degrees of Freedom d_1	d_2
1	1	111
2	2	110
3	3	109

Example (3):

Step number	$F_{.95}(1, d_2-1)$
1	3.96
2	3.96
3	3.98

PROCEDURE

(4) Look up, for the variable entered at each step, the F-to-enter or -remove value from the summary table. Compare the F-to-enter or -remove value with the $F_{1-\alpha}(1,d_2-1)$ value obtained in (3) at each step: If the F-to-enter or-remove value is greater than $F_{1-\alpha}(1,d_2-1)$, then the variable entered in that step should be included in the discriminant function; otherwise, the variable is excluded from the discriminant function.

(5) Look up the coefficient table for the classification function at step k in the printout. Here, k is the number of variables in the best subset obtained in (4).

Step number k

⋮

Classification functions

EXAMPLE

(4) Look up, for the variable entered at each step, the F-to-enter or-remove value from the summary table.

Step Number	Variable Entered	F-to-Enter or/-Remove
1	X_1	131.41
2	X_3	17.35
3	X_2	2.55

Compare with the table in (3). At step 1, the F-to-enter for X_1 is greater than 3.96; thus, X_1 is included in the discriminant function. At step 2, the F-to-enter for X_3 is greater than 3.96; thus, X_3 is included. At step 3, the F-to-enter for X_2 is less than 3.98; thus, X_2 is excluded. Therefore, the best set consists of X_1 and X_3.

(5) Look up the coefficient table for the classification function at step 2 in the printout.

Step number 2

⋮

Classification functions

Variable	Group = A	B
X_{K1}	a_{11}	a_{12}
X_{K2}	a_{21}	a_{22}
.	.	.
.	.	.
.	.	.
X_{KK}	a_{K1}	a_{K2}
Constant	C_1	C_2

(6) Compute the Fisher discriminant function from the coefficient table in (5) according to the formula

$$Z = (a_{11}-a_{12})X_{K1}+(a_{21}-a_{22})X_{K2} + \dots +(a_{K1}-a_{K2})X_{KK}+C_1-C_2$$

(7) Classify an X into group A if

$$Z = (a_{11}-a_{12})X_{K1}+ \dots +(a_{K1}-a_{K2})X_{KK} + C_1-C_2 \geq 0$$

Here, $X_{K1}, \dots, X_{KK}$ significantly contribute to the discrimination of the two groups at the α significance level.

Variable	Group = 1	2
X_1	0.690	0.544
X_3	211.621	199.016
Constant	-213.956	-182.326

(6) Compute the Fisher discriminant function from the coefficient table in (5) by

$$Z = (0.690-0.544)X_1 + (211.621-199.016)X_3 + (-213.956 + 182.326)$$
$$= 0.146\ X_1 + 12.605\ X_3 - 31.63$$

(7) Classify a car with emission vector (NO_x,CO,HC) into Category I, if $0.146 \log (NO_x) + 12.605(HC) -31.63 \geq 0$. Here, $\log(NO_x)$ and HC significantly contribute to the discrimination of the two categories at the $\alpha = 0.05$ level.

Procedure for computing the estimated Mahalanobis distance D_q^2 for each step q (required in Stage 4).

PROCEDURE

(1) For each step q, look up the value of approximate F-statistics (or equivalently, Rao's V in SPSS) from the summary table and copy it.

(2) For each step q, compute the Mahalanobis distance D_q^2 by using the relationship

$$D_q^2 = \frac{q(N_1+N_2)(N_1+N_2-2)}{N_1N_2(N_1+N_2-q-1)} \quad F,$$

where N_1 is the sample size of group A and N_2 is the sample size of group B, and F is the value found in (1).

EXAMPLE

(1) Look up the value of approximate F-statistics from the summary table.

Step Number	Approximate F-Statistics
1	131.41
2	84.06
3	63.58

(2) For step 2, compute D_2^2 by

$$D_2^2 = \frac{2(70+43)(70+43-2)}{70\text{x}43(70+43-2-1)} \times 84.06$$

$$= 6.31.$$

Friedman's Recursive Partitioning Algorithm for Nonparametric Classification. If the underlying distributions for the short-test emission data are not normal, the classification will be performed using Friedman's recursive algorithm. This algorithm is the newest development in the state of the art. A complete program write-up and the user's manual are available at TSC. Typically, the use of the algorithm will result in a tree-structured decision rule: Starting at the root, the test vector is directed down the tree until it arrives at a terminal node and is assigned to the class with the highest representation at the node.

Figure 2.10 illustrates a hypothetical decision tree, where X_1, X_2 and X_3 denote the short-test NO_x, CO and HC values, respectively, and α's are the cutoff points. Starting at the top node, if a car's short-test emission NO_x value is less than α_1, descend to the right node and examine its short-test HC value. If this value is greater than α_3, descend to the left terminal node and classify the car into Category II as dictated by the class labels at that node. If, on the other hand, a car's short-test NO_x value is greater than α_1, descend to the left node; and if the short-test CO value is less than α_2, classify the car into Category I, according to the class label.

Structural relationships among the three variables, insofar as classification is concerned, can be accessed immediately from the tree structure. For example, at the top root, the short-test NO_x clearly is the most influential variable in the classification. However, the short-test CO and HC values are independent of each other in the decision

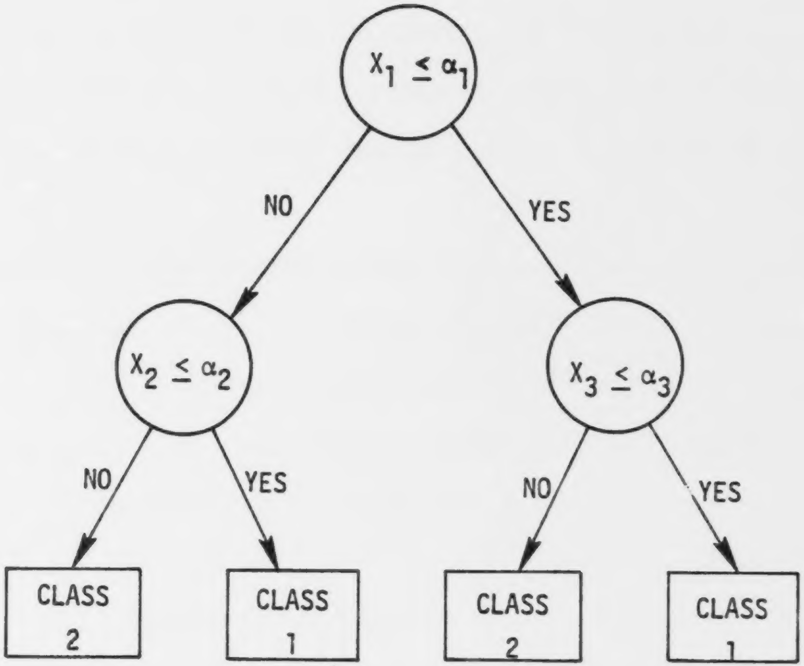

$X_1 \leq \alpha_1$
NO
YES
$X_2 \leq \alpha_2$
$X_3 \leq \alpha_3$
NO
YES
NO
YES
CLASS 2
CLASS 1
CLASS 2
CLASS 1

Figure 2.10 Hypothetical decision tree.

process; i.e., when a car's NO_x exceeds α_1, its CO value is then the dominant factor in determining the class. When the car's NO_x is lower than α_1, its HC value will be the dominant factor in discriminating the two classes.

In general, the advantages of a tree-structured method include:

1. Being able to fully exploit variables that are important only under limited conditions, because variable use is determined by previous structure values.
2. Not having to calculate a variable if classification is possible without that variable.
3. An algorithm that readily permits interpretation.
4. An efficient decision rule (the tree): Few calculations are required to classify a sample.

The construction of such a tree proceeds as follows: At the top, put the root of the tree, representing the entire sample. Then form splits, the two descendants of the root or of each nonterminal node representing the two subsamples defined by the truth or falsity of the inequality

$$x_k \leq \alpha .$$

Choosing the particular variable x_k and the cutoff point α is done by searching through all the allowable splits for all the variables and picking the split of that particular variable which results in the greatest decrease in uncertainty insofar as classification is concerned. The degree of uncertainty is measured by the Kolmogorov-Smirnov distance between the empirical marginal distributions of the two classes [Friedman 1977]. Terminal conditions are needed to keep the program from splitting down to a point of insignificance. A typical terminal condition has the form: Do not split a node if either a) the node contains less than m points, where m is the threshold value chosen by the user, or b) the percentage of the majority class at the node is > 90%.

EXAMPLE: The Friedman's algorithm is applied to a real data set from the FTP malfunction study. The objective of the study is to predict the pass or fail of the FTP, based on various short-test emission measurements. There are 619 car tests involved in the study. For each car test, the data consists of a record of its pass or fail of the FTP, along with measurements of seven variables: idle CO, idle speed, choke adjustment, vacuum leak, timing, EGR, and manufacturer.

Before applying the Friedman's algorithm, the data are randomly split into two groups. One group of 397 car tests, called the Learn Mode, is used to design the classification tree; another group of 222 car tests, called the Test Mode, is used to assess the accuracy, or the probability of error, of the tree designed.

The details of the computer subroutine, along with the user's manual, are given in Appendix B. Output examples are given in Table 2.3, and an interpretation of the output follows.

TABLE 2.3 OUTPUT EXAMPLE (Friedman's Algorithm)

LEARN MODE

Node	Variable	Cutpoint	From	To left	To right	PASS	FAIL	Rule	Class	P(OR)%
1	IDLE CO	0.300	0	2	3	199	198	0	FAIL	50.00
2	EGR	0.00	1	4	5	181	112	0	PASS	61.66
3	IDLE CO	4.20	1	6	7	18	86	0	FAIL	82.76
4			2	Terminal		179	93	3	PASS	65.70
5			2	Terminal		2	19	3	FAIL	90.52
6	IDLE SPD	0.00	3	12	13	16	42	0	FAIL	72.51
7			3	Terminal		2	44	2	FAIL	95.67
12			6	Terminal		15	24	3	FAIL	61.66
13	CHOKE AD	-3.00	6	20	21	1	18	0	FAIL	94.76
20			13	Terminal		1	0	1	PASS	100.00
21			13	Terminal		0	18	2	FAIL	100.00

Classification was made using the prior probabilities of 0.5012 for PASS FTP and 0.4987 for FAIL FTP.

TABLE 2.3--Cont'd

```
MISCLASSIFICATION Output for LEARN MODE

                               Classified
                         PASS FTP     FAIL FTP

True        PASS FTP       180           19
Classes     FAIL FTP        93          105

Percentage correct = 71.7884

Percentage wrong   = 28.21159


MISCLASSIFICATION Output for TEST MODE

                               Classified
                         PASS FTP     FAIL FTP

True        PASS FTP        96           17
Classes     FAIL FTP        50           59

Percentage correct = 69.8198

Percentage wrong   = 30.1801

Expected percentage correct = 72.0787
```

INTERPRETATION OF THE OUTPUT

PROCEDURE	EXAMPLE
(1) Construct a tree diagram from the summary table under Learn Mode. Starting at node 1, the variable name, x_k, appears in the second column. The cutoff point, α, given in the third column, represents a split of the node into a left node (5th column) and a right node (6th column), according to the truth or falsity of the inequality $x_k \leq \alpha$. For each of the descendant nodes, look for the corresponding splitting variable, the cutpoint, the left and the right nodes. If there is a "terminal" appearing in the To left and To right columns, the splitting stops and that node is considered the terminal node.	(1) Construct a tree diagram from the summary table shown in the Learn Mode of the Output Example. At node 1, the splitting occurs in IDLE CO, with cutpoint 0.300. Thus, a sample point with IDLE CO less than or equal to 0.300 will be directed to the left node (2), while any point with IDLE CO greater than 0.300 will be assigned to the right node (3). At node 2, the splitting occurs in EGR, with cutpoint 0.00; the two descendants are nodes 4 and 5. Since "terminal" occurs at nodes 4 and 5, the splitting stops. Similarly, at node 3, the splitting occurs in IDLE CO with cutpoint 4.20, and the two descendants are nodes 6 and 7. A complete tree structure generated from the summary table is shown in Figure 2.11.

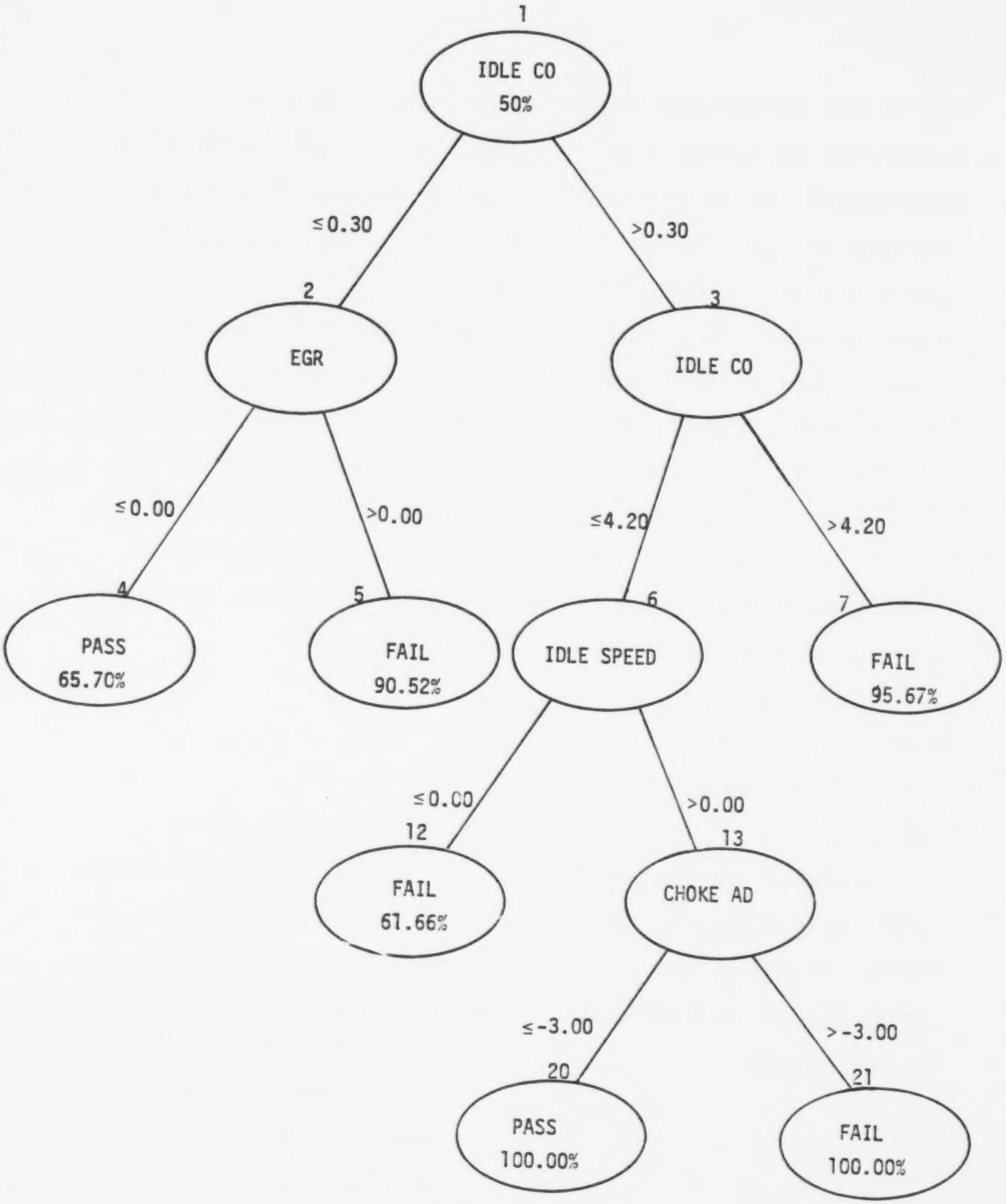

1
IDLE CO
50%
≤0.30
>0.30
2
EGR
3
IDLE CO
≤0.00
>0.00
≤4.20
>4.20
4
PASS
65.70%
5
FAIL
90.52%
6
IDLE SPEED
7
FAIL
95.67%
≤0.00
>0.00
12
FAIL
61.66%
13
CHOKE AD
≤-3.00
>-3.00
20
PASS
100.00%
21
FAIL
100.00%

Figure 2.11 Tree structure of the Output Example.

(2) At each terminal node, the CLASS column gives the decision label; a point which follows the tree and falls into a terminal node with class labeling FAIL will be classified as failing the FTP. The PASS column gives the number of sample points at a node that pass the FTP. The FAIL column gives the number of sample points that fail. The P(OR)% column from Table 2.3 gives the probability that a point falling into the node will be correctly classified. In other words, P(OR)% provides a confidence level for the classification at a given node.

(3) The table in the MISCLASSIFICATION Output in the Learn Mode provides one with an overall impression of the performance of the decision rule.

(2) The class label at node 5 is FAIL. Thus, any point which follows the tree branches IDLE CO $\leq$ 0.30 and EGR > 0.00 will be classified as failing the FTP. The number shown in the P(OR)% column is 90.52%. This means that, at node 5, there is a 90% confidence level in making the decision as FAIL category. At node 4, this number is 65.70%, and therefore one is less confident in making the decision as PASS category.

(3) In the MISCLASSIFICATION Output, the following table provides one with an overall impression of the performance of the decision rule:

	Classified	
	PASS FTP	FAIL FTP
PASS FTP	180	19
FAIL FTP	93	105

Percentage correct = 71.7884
Percentage wrong = 28.21159

(4) The MISCLASSIFICATION output for the TEST MODE provides one with an impression of the accuracy of the decision rule, based on data in the LEARN MODE. An estimate of the probability of misclassification is given by the PERCENTAGE WRONG. A much more accurate estimate can be obtained by using the "leave-m-out" technique on the sample; a small number m of vectors (usually 10% to 20% of the total sample size) are formed in the TEST MODE and the remainder are put in the LEARN MODE to design the tree. Record the PERCENTAGE WRONG in the TEST MODE. This procedure is repeated 10 to 20 times, each time with a different set of vectors in the TEST MODE. The averaged PERCENTAGE WRONG value over all of these trials is an accurate estimate of the probability of misclassification for the decision tree.

(4) In the MISCLASSIFICATION output for the TEST MODE, find the PERCENTAGE WRONG = 30.1801. This is one estimate of the probability of misclassification. The "leave-m-out" method will provide one with a much more accurate estimate.

2.1.4 Stage 4. Evaluation of the ST/FTP Association

The classification method of Stage 3 is employed to establish a prediction scheme for the FTP categorization based on a car's short-test emission values. Such a scheme generally produces two kinds of errors, false positive errors and false negative errors. Let E_0 and E_1 be the probabilities of producing a false positive error and of producing a false negative error, respectively, and let P be the probability that an observation comes from Category I. (Thus 1-P is the probability that an observation comes from Category II). Then an optimal classification scheme is the one which achieves the minimum expected probability of misclassification, E*, where $E^* = (1-P)E_0 + PE_1$. Under the joint normality assumption, the discriminant analysis scheme is optimal; under the non-normality assumption, the Friedman's scheme tends to be "more" optimal than the discriminant scheme. The degree of ST/FTP association can now be defined as that value of E* achieved by using the optimal classification scheme. The smaller the value E*, the higher the ST/FTP association and vice versa. This E* can be best utilized in comparing the classification results for various portions of the data, which will be the main subject of Stage 5 of the analysis.

In practice, however, the true magnitudes of P, E_0 and E_1 are unknown, and therefore have to be estimated from the data. A common estimate of P is the number $N_1/(N_1+N_2)$, where N_1 and N_2 are the sample sizes for Category I and Category II, respectively. The estimations of E_0 and E_1 are much more involved, since the methods depend on the particular classification algorithm being used. The details of estimating errors, along with the details of obtaining the 95% confidence intervals, which provide one with an impression of the reliability of the estimates of errors, are given below.

Estimating the error rates and 95% confidence intervals when the discriminant analysis algorithm is used. There are two methods for estimating the error rates and one method for constructing the 95% confidence interval, as follows:

Procedure 1 (biased estimates) -- If M_1 is the number of observations from Category I that are classified into Category II, and M_2 is the number of observations from Category II that are classified into Category I, then estimate E_0 by M_2/N_2 and E_1 by M_1/N_1. These two estimates are readily available in the printout of the stepwise discriminant analysis subroutine described in Stage 3 (p.2-44 and p.2-47).

PROCEDURE

(1) Look up the classification matrix in the printout of the stepwise discriminant analysis (p.2-44).

Classification Matrix

Group	Percentage Correct	Number of cases classified into group A	B
A	p	a	b
B	q	c	d

EXAMPLE

(1) Look up the classification matrix in the output example (BMDP7M, p.2-44).

Classification Matrix

Group	Percentage Correct	Number of cases classified into group 1	2
1	95.7	67	3
2	83.7	7	36
TOTAL	89.7	74	39

(2) Estimate E_0 by (100-q)% and E_1 by (100-p)%.	(2) Estimate E_0 by 100-83.7=16.3% and E_1 by 100-95.7=4.3%.

COMMENT: Since these two estimates are known to be biased (i.e., they underestimate the errors), they can only be used as rough estimates.

Procedure 2 (unbiased estimates)--A satisfactory method of estimating the probabilities of errors has been provided by Okamoto [1963]. The method involves the computation of the estimated sample Mahalanobis distance, which is a byproduct of every stepwise discriminant subroutine described in Stage 3. The details of the procedure, along with the numerical example of Stage 3 (p.2-51), are given below.

PROCEDURE	EXAMPLE
(1) Compute the estimated Mahalanobis distance D^2 for step k in the printout of the stepwise discriminant analysis, where k is the size of the best subset of variables chosen by the program. (See Stage 3, p.2-51.)	(1) Assume that in the printout of the stepwise discriminant analysis, the size for the best subset is 3. Thus, according to the procedure for computing the Mahalanobis distance (p.2-51), the D^2 for Step (3), following, is D^2=1.2 (or D=1.09).
(2) Estimate P by $N_1/(N_1+N_2)$.	(2) Estimate P by 70/(70+43)=2/3.

(3) Compute $c = \left\{ \log[\frac{(1-P)}{P}] - \frac{D^2}{2} \right\} / D$	(3) Compute $c=[\log(.5)-0.6]/1.09=-1.186$
(4) Compute $a_1 = 3c-c^3+K(D-c)$ $a_2 = 2D+3c-c(D+c)^2-k(D+c)$ $a_3 = 2(D+3c)-c(D+2c)^2-2k(D+3c)$	(4) Compute $a_1=4.938$ $a_2=0.297$ $a_3=8.351$
(5) Let Φ and ϕ denote the cumulative distribution function and the density function, respectively, of the standard normal distribution. Look up $\Phi(c)$ from the normal table (Table A-3), and compute $\phi(c)=\frac{1}{\sqrt{2\pi}} e^{-c^2/2}$.	(5) Look up $\Phi(-1.186)$ from the normal table and get $\Phi(-1.186)=1-\Phi(1.186)$ $=1-0.8810$ $=0.1190$ $\phi(-1.186)=0.2792$.
(6) Let $N=N_1+N_2$. Set $E_1(c,D)=\Phi(c)+(\phi(c)/N)(\frac{a_1}{2PD^2}+\frac{a_2}{2(1-P)D^2}+\frac{a_3}{4})$ $E_0(c,D)=\Phi(c)-(\phi(c)/N)(\frac{a_2}{2PD^2}+\frac{a_1}{2(1-P)D^2}+\frac{a_3}{4})$	(6) $N=43+70=113$. Set $E_1=.2412$ $E_0=.0601$
(7) Set $E^*=PE_1(c,D)+(1-P)E_0(c,D)$ This is an estimate of the expected probability of misclassification.	(7) Set $E^*=\frac{2}{3}(.2412)+\frac{1}{3}(.0601)=0.18$ This is an estimate of the expected probability of misclassification.

Procedure 3 (95% confidence interval) -- A method for constructing an accurate 95% confidence interval around the individual probability of errors has been developed by McLachlan [1975]. The details of the computational procedures are described below with notations and numerical examples of Procedure 2 carried over.

PROCEDURE	EXAMPLE
(1) Select a fixed 100(1-α)% level-of-confidence interval.	(1) Select 95% level-of-confidence interval, i.e., set $\alpha=0.05$.
(2) Copy the values of k, P, D^2, c, N, $E_1(c,D)$ and $E_0(c,D)$ obtained from the aforementioned Procedure 2 (unbiased estimate of error rates).	(2) From the previous example (see Procedure 2), one obtains $k=3$, $P=\frac{2}{3}$, $D^2=1.2$, $c=-1.186$, $N=113$, $E_1=0.2412$, and $E_0=0.0601$.
(3) Compute $b_1=D^2+4(3k-4)+(k^2-4k+5)(16/D^2)$ $b_2=(D^2-2k)/8$ $b_3=D^4+2(11k-16)D^2+8(5k-4)$ $b_4=D^4-2(k+4)D^2+8k$ $b_5=2D^6+16(2k-5)D^4-32(4k-13)D^2$	(3) Compute b_i, $i=1,\ldots,5$. According to the formulas given on the left side, one gets $b_1=47.86$ $b_2=0.6$ $b_3=130.8736$ $b_4=9.2736$ $b_5=77.5495$
(4) Compute $\phi(c) = \frac{1}{\sqrt{2\pi}} e^{-c^2/2}$	(4) $\phi(-1.186)=0.2792$

(5) Set

$$\sigma_1^2=(\phi(c))^2[(\frac{1}{P}+\frac{D^2}{8})/N+b_1/(4PN)^2$$
$$+b_2/(P(1-P)N^2)$$
$$+b_3/(64PN^2)+b_4/(64(1-P)N^2)$$
$$+b_5/(32N)^2]$$

$$\sigma_0^2=(\phi(c))^2[(\frac{1}{1-P}+\frac{D^2}{8})/N+b_1/(4(1-P)N)^2$$
$$+b_2/(P(1-P)N^2)+b_3/(64(1-P)N^2)$$
$$+b_4/(64PN^2)+b_5/(32N)^2]$$

(5) Compute σ_1^2, σ_0^2. According to the formulas given on the left, one gets

$$\sigma_1^2 = 0.0003$$

$$\sigma_0^2 = 0.021$$

(6) Let Z_α be the 100α percentile of the standard normal distribution. For the α given in (1), look up $Z_{\alpha/2}$ from the normal table (Table A-3). The approximate $100(1-\alpha)\%$ confidence interval for the probability of false negative error is $E_1 \pm \sigma_1 \cdot Z_{\alpha/2}$; while the approximate $100(1-\alpha)\%$ confidence interval for probability of false positive error is

$$E_0 \pm \sigma_0 \cdot Z_{\alpha/2}$$

(6) Since $\alpha=.05$, look up $Z_{.025}$ from the normal table and get $Z_{.025}=1.96$. The approximate 95% confidence interval for the probability of false negative error is

(0.2411, 0.2413).

The approximate 95% confidence interval for the possibility of false positive error is

(0.0523, 0.0679).

Interpretation of the 95% confidence interval. The preceding method constructs 95% confidence intervals for individual probabilities of error. The interpretation for those two confidence intervals can be explained as follows: Let I_i be the 95%

confidence interval for estimating E_i, where i=0 or 1. If the experiments can be repeated a large number of times, then, each time the classification algorithm is performed and the estimated E_0 and E_1 are computed, roughly 95 out of 100 times, the estimated E_0 will fall into the interval I_0 and,similarly,95 out of 100 times, the estimated E_1 will fall into the interval I_1. But this is not the same as saying 95 out of 100 times, E_0 will lie in I_0 and E_1 will lie in I_1 "simultaneously". In fact, a joint 95% confidence region for E_1 and E_0 can be easily derived from an inequality found by Bonferoni,* which roughly states that the 97.5% confidence intervals for each of the two probabilities of errors ensure a "joint" 95% confidence region for both probabilities. Thus, using the previous method, construct 97.5% confidence intervals I_0 and I_1 for E_0 and E_1,respectively. Then E_0 will lie in I_0 and E_1 will lie in I_1 simultaneously with at least 95% confidence.

---

*Bonferoni inequality: Let I_i be the $100(1-\alpha)\%$ confidence interval of E_i when i=0,1; then

$$\text{Prob }(E_0 \text{ is in } I_0 \text{ and } E_1 \text{ is in } I_1) \geq 1-\text{Prob}(E_0 \text{ is not in } I_0) - \text{Prob}(E_1 \text{ is not in } I_1) = 100\% - 100\alpha\% - 100\alpha\% = 100(1-2\alpha)\%.$$

2.1.5 Stage 5. Testing Hypotheses

The in-use vehicles tested in the Portland study were split into two fleets, Elements I and II [Hamilton Test Systems 1977]. Element I vehicles were recruited exclusively from the Portland region; Element II vehicles were recruited jointly from Portland and Eugene, Oregon. Each fleet is defined by several vehicle characteristics. The following questions relating to the fleets are of immediate concern:

- How would one evaluate the ST/FTP association, given the two test fleets' emission data?
- Does the ST/FTP association depend on vehicle characteristics used to define the fleets? If so, can the degree of association be improved upon by taking these characteristics into account?

These questions can be answered through statistical hypothesis testing procedures. Owing to the nature of the problem and the particular sampling designs of the test fleets, a multi-contingency table approach (an approach entirely different from the one mentioned in "Short Test Correlation Analysis" [Aerospace Corporation 1976]) is appropriate. Since the Element I and Element II vehicle fleets were recruited by using different sampling methods, it is more convenient to discuss separately the hypothesis testing problem for each fleet.

2.1.5.1 Element I Fleet. The Element I vehicle fleet for the Portland Study comprises 2,400 1975 through 1977 model-year cars, recruited exclusively from the Portland region. A stratified sampling method was used to recruit the cars. The Portland vehicle population was divided into 60 vehicle groups, according to manufacturer, model year, engine family, transmission, and test inertia weight in an attempt to identify different emission control technologies. A sample of 40 vehicles was taken for each vehicle group, for a total of 2,400 cars.

As demonstrated above, the ST/FTP association can be evaluated using a classification algorithm (Stages 3 and 4) for a given set of FTP standards. Since the FTP emission standards for 1975-76 model-year cars are not as stringent as those for 1977 cars, it is best to apply the classification scheme separately to 1975-76 cars and 1977 cars and then to study the resulting classifications.

Using 1975-76 cars as examples, the procedures for testing whether the classification results depend on various vehicle characteristics and for fine-tuning the classification schemes are described below.

1.A) Combining Vehicle Group Cells.

Among the 60 cells for the Element I fleet sampling design, 44 cells are occupied completely or partly by the 1975-76 model-year cars. Following the procedures described above in Stages 3 and 4, a classification rule and expected probability of misclassification, E*, based on the ST/FTP emission data of these 1975-76 cars, can be obtained. Here, E* measures the overall performance of the classification rule. The misclassification rate of the individual cells indicates how well the rule works for each particular class of cars. An unusually large misclassification rate in one cell indicates the inappropriateness of the rule for that cell.

The following procedures can be used to form blocks of cells, so that each block comprises one or more cells with similar classification characteristics. Thus, the appropriateness of the rule when applied to a particular block can be judged from the common misclassification rate.

Method

(a) Construct a two-way table (table of counts).

There are two variables: classification (correctly classified and misclassified), and the vehicle-group cell number (44 categories for 75-76 model-year cars.)

Let $\{X_{1i}\}$ and $\{X_{0i}\}$ be two 1x44 tables, where X_{1i} and X_{0i} are the number of misclassified cars and the number of correctly classified cars, respectively, for the i^{th} cell. Let $n_i = X_{1i} + X_{0i}$. Then $r_i = X_{1i}/n_i$ gives the misclassification rate fc · the i^{th} cell.

NOTE: For 75-76 model cars, there are 35 cells with sample sizes 40.

(b) Order the 44 rates from smallest to largest, denoting by $r_{(i)}$ the i^{th} smallest rates $(i=1,\ldots,44)$,

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(44)}.$$

(c) Group those rates which are most alike in magnitude into one class, either by eyeball inspection or by the following more rigorous method: First, test the hypothesis H_0: $r_{(1)} = r_{(2)}$, using the chi-square test to be described below. If the test result is to accept the hypothesis, combine the two cells into one block and continue to test H_0: $r_{(1)} = r_{(2)} = r_{(3)}$. Otherwise, consider the first cell as a single block and proceed to test H_0: $r_{(2)} = r_{(3)}$. Continue the process until $r_{(k)}$ is reached.

The chi-square test used in combining cells is as follows:

(i) Suppose one wants to test whether $r_1,\ldots,r_k$ are the same in magnitude, i.e., test the hypothesis

$$H_0: r_1=r_2=\ldots=r_k.$$

(ii) Compute

$$X_{1+} = \sum_{i=1}^{k} X_{1i} \text{ and } X_{0+} = \sum_{i=1}^{k} X_{0i}.$$

(iii) Compute the statistics

$$\chi^2 = \left[\sum_{i=1}^{k} (X_{1i} - \frac{X_{1+}}{k})^2/(X_{1+}/k) + (X_{0i} - \frac{X_{0+}}{k})^2/(X_{0+}/k)\right].$$

(iv) Under the hypothesis, χ^2 has an approximate chi-square distribution with k-1 degrees of freedom. Look up the p value of χ^2 from the $\chi^2(k-1)$ frequency table, i.e., the area to the right of the χ^2 value under the $\chi^2(k-1)$ curve (Table A-6, Appendix A). If p is greater than 0.05, then accept the hypothesis; otherwise, reject the hypothesis.

1.B) Testing Whether the ST/FTP Association Depends on Various Vehicle Characteristics.

The classification rules designed according to Stage 3 or 1.A enable one to assess the ST/FTP association. But this association may depend on various engine parameters, e.g., manufacturer, model year, engine family, and CID. In other words, some significant interaction effect between the classification and a specific engine parameter may exist. A valid testing hypothesis procedure can be used to identify such interactions.

In principle, the procedure will test the null hypothesis by fitting a model with the selected interaction effect set equal to zero, then will fit a second model in which this interaction effect is included. The difference between the goodness-of-fit of the two models becomes the test statistic of the null hypothesis, provided that the second model fits the data well.

The details of the testing hypothesis are described below.

1.B.1 Select the primary variables.

In deciding what variables should be included in the testing of interaction effects, three variables at the most are considered each time, for ease in interpretation.

EXAMPLE: Suppose there are I categories for variable 1, J categories for variable 2, and K categories for variable 3. The primary variables are classification (2 categories). MAKE (combined into 5 categories for 75-76 model-year cars), and CID (combined into 8 categories).

1.B.2 Construct the contingency table (table of counts).

Let $\{X_{1jk}\}$ and $\{X_{0jk}\}$ be two JxK tables ($j=1,\ldots,J$, $k=1,\ldots,K$), where X_{1jk} and X_{0jk} are the number of misclassified cars and the number of correctly classified cars, respectively, in the jk-cell. Thus, $n_{jk}=X_{1jk}+X_{0jk}$ is the total number of sample cars in the jk-cell.

EXAMPLE: Let $\{X_{1jk}\}$ and $\{X_{0jk}\}$ be two 5x8 tables, where X_{1jk} and X_{0jk} are the number of misclassified cars and the number of correctly classified cars, respectively, in the j^{th} MAKE and k^{th} CID cells. These two tables can be combined into a 2x5x8 multicontingency table, i.e., into table $\{X_{ijk}\}$, where $i=1,0$.

1.B.3 Construct a linear model in the natural logarithms of the cell counts;[*] the linear model is analogous to an analysis of variance (ANOVA) model. The advantages of such a model have been pointed out in Bishop, et al. [1975].

Let $\{m_{ijk}\}$ be the expected value of the observed frequencies $\{X_{ijk}\}$. Write the full model

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(i)}+U_{3(k)}+U_{12(ij)}+U_{13(ik)}+U_{23(jk)}+U_{123(ijk)},$$

where $i=0,1$; $j=1,\ldots,J$; $k=1,\ldots,K$.

The interpretation of the U-terms is as follows:

U is the overall mean

U_1 measures the main effect of the classification variable

U_2 measures the main effect of the MAKE variable

U_3 measures the main effect of the CID variable

U_{12} measures the interaction between classification and MAKE

U_{13} measures the interaction between classification and CID

U_{23} measures the interaction between MAKE and CID

U_{123} measures the difference between MAKE-CID interaction effects attributable to the classification variable.

Various models can be derived from the above full model by setting particular U-terms equal to zero:

---

[*] Grizzle et al. [1969] have proposed the use of a weighted least-square technique for analyzing categorical data. This technique has the advantage of giving the user more latitude in choosing the model and testing hypothesis. But the disadvantages are that 1) it requires the user to have a more sophisticated background in statistics to actually carry out the analysis and 2) technical difficulties arise when there are sporadic zero cells.

Model (a): Test $U_{123}=0$. Thus, the model is

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}+U_{13(ik)}+U_{23(jk)},$$

which states that the table can be described with constant two-factor effects; for example, constant classification-MAKE, classification-CID, and MAKE-CID effects.

Model (b): Test $U_{123}=0$ and one of the two-factor effects is missing. There are three versions of this model:

Model (b.1) Selecting $U_{13}=0$ gives

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}+U_{23(jk)}.$$

This model states that the classification and CID are mutually independent for every MAKE. A direct consequence of the model is that CID can be completely deleted from the analysis.

Model (b.2) Selecting $U_{12}=0$ gives

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{13(ik)}+U_{23(jk)}.$$

This model states that classification and MAKE are mutually independent for every CID level. Thus, MAKE can be completely deleted from the analysis.

Model (b.3) Selecting $U_{23}=0$ gives

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}+U_{13(ik)}.$$

This model states that MAKE and CID are mutually independent for each of the two classification levels.

Model (c): Test $U_{123}=0$, and two of the two-factor effects are zero. There are also three versions of this model:

Model (c.1) Selecting $U_{123}=U_{13}=U_{23}=0$ gives

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{12(ij)}.$$

CID is now completely independent of classification and MAKE; classification status and MAKE are associated.

Model (c.2) Selecting $U_{123}=U_{13}=U_{12}=0$ gives

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{23(jk)}.$$

Classification is completely independent of CID and MAKE.

Model (c.3) Selecting $U_{123}=U_{12}=U_{23}=0$ gives

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)}+U_{13(ik)}.$$

MAKE is completely independent of classification and CID; classification and CID are associated.

Model (d): Test $U_{123}=U_{12}=U_{13}=U_{23}=0$. This gives the model

$$\log(m_{ijk})=U+U_{1(i)}+U_{2(j)}+U_{3(k)},$$

which states that all three variables are completely independent.

Let $\{X_{ijk}\}$ be the observed frequencies of counts. Then various tables of "sums" can be derived. Such tables are the necessary ingredients for estimating the expected frequencies of counts m_{ijk}. Let C_{12} be the IxJ table of sums $\{X_{ij+}\}$, where $X_{ij+} = \sum_{i=1}^{K} x_{ijk}$. Similarly, let $C_{13}=\{X_{i+k}\}$ and $C_{23}=\{X_{+jk}\}$. Let C_1 be the Ix1 table of sums $\{X_{i++}\}$, where $X_{i++} = \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk}$. Similarly, let $C_2=\{X_{+j+}\}$ and $C_3=\{X_{++k}\}$.

Table 2.4, which will be used later, gives a complete account of the models in three dimensions.

1.B.4 Procedure for fitting a model.

Given the observed frequencies of counts $\{X_{ijk}\}$, the estimates $\hat{m}_{ijk}$ of the expected value m_{ijk} under various model assumptions listed in 1.B.3 can be calculated with the use of algorithm AS-51, commonly called the "Iterative Proportional Fitting Algorithm" and developed by Haberman [1972]. Complete details of the algorithm, including program card preparation and standard FORTRAN statements are given in Appendix C. An experienced programmer should have no difficulty in implementing the program by himself/herself.

TABLE 2.4 COMPLETE MODELS IN THREE DIMENSIONS

Model Type	Absent Term	Sufficient Configurations	Degree Freedom
(a)	U_{123}	$C_{12}\ C_{13}\ C_{23}$	(I-1)(J-1)(K-1)
(b)	$U_{12}\ U_{123}$	$C_{13}\ C_{23}$	(I-1)(J-1)K
	$U_{13}\ U_{123}$	$C_{12}\ C_{23}$	(I-1)J(K-1)
	$U_{23}\ U_{123}$	$C_{12}\ C_{13}$	I(J-1)(K-1)
(c)	$U_{12}\ U_{13}\ U_{123}$	$C_{23}\ C_{1}$	(I-1)(JK-1)
	$U_{12}\ U_{23}\ U_{123}$	$C_{13}\ C_{2}$	(J-1)(IK-1)
	$U_{13}\ U_{23}\ U_{123}$	$C_{12}\ C_{3}$	(K-1)(IJ-1)
(d)	$U_{12}\ U_{13}\ U_{23}\ U_{123}$	$C_{1}\ C_{2}\ C_{3}$	IJK-(I+J+K)+2

SPECIAL PROGRAM CARD PREPARATION

AS-51 -- The details of card preparation are given in Appendix C. Those input statements which require special instructions and those output statements which deserve special attention are listed below.

INPUT STATEMENT	EXAMPLE
(1) NVAR=the number of variables in the table.	(1) NVAR = 3
(2) DIM Integer array (NVAR), the number of categories in each variable of the table.	(2) DIM(1) =2, DIM(2) = 5 DIM(3) = 8
(3) NTAB = the number of cells in the table.	(3) NTAB=80
(4) TABLE Real array (NTAB), the table to be fit.	(4) TABLE (1)=X_{111}, TABLE(2)=X_{211}, TABLE(3)=X_{121},TABLE(4)=X_{221}, ..., TABLE(80)=X_{258}
(5) FIT Real array (NTAB),the fitted table.	(5) FIT(1)=1, FIT(2)=1, ... , FIT(80)=1
(6) MAXDEV = 0.25 (or 0.05)	(6) MAXDEV = 0.25 (or 0.05)

(The following statements depend on the model type to be fitted.)

(7) NCON = the number of sufficient configurations for a given model. Here,the sufficient configurations can be found in the table of three-dimensional models, Table 2.4.

(7)

	MODEL			
	1	2 (U_{13}=0)	3 (U_{12}=U_{13}=0)	4
NCON=	3	2	2	3

(8) CONFIG Integer array (NVAR,NCON), the sets indicating the sufficient configurations to be fit. For a given model, the sufficient configuration can be found in Table 2.4.	(8) CONFIG(1,1)= 1 1 0 1 CONFIG(2,1)= 2 2 2 0 CONFIG(3,1)= 0 0 3 0 CONFIG(1,2)= 1 0 1 0 CONFIG(2,2)= 0 2 0 2 CONFIG(3,2)= 3 3 0 0 CONFIG(1,3)= 0 0 0 0 CONFIG(2,3)= 2 0 0 0 CONFIG(3,3)= 3 0 0 3
(9) MAXIT Integer, the maximum permissible number of iterations.	(9) MAXIT = 10 1 1 1
(10) NMAR Integer, the maximal dimension of the sufficient configurations for a given model. This dimension can be found from the configurations listed in Table 2.4.	(10) NMAR = 40 40 40 40
(11) NU Integer, the dimension of a work area, U, used to store fitted configurations.	(11) U = 40. 40 40 40.

OUTPUT STATEMENT

(1) The maximum likelihood estimates for the cell counts are given in the following FIT table:

FIT(1)=$\hat{m}_{111}$, FIT(2)=$\hat{m}_{211}$, FIT(3)=$\hat{m}_{121}$, FIT(4)=$\hat{m}_{221}$, ...,FIT(80)=$\hat{m}_{258}$

(2) The number of the last iteration will be shown in NLAST.

(3) In case there is any error in the program, an error message is given by one of the following IFAULT statements:

If IFAULT=0, then no error was detected

If IFAULT=1, then CONFIG contains an error

If IFAULT=2, then MARG or U is too small

If IFAULT=3, then the algorithm did not converge to the desired accuracy within MAXIT iteration (change MAXIT to a larger value)

If IFAULT=4, then DIM, TABLE, FIT or NVAR is erroneously specified.

1.B.5 Goodness-of-fit for models computed by using AS-51 algorithm.

In order to measure how well the selected model fits the data, the likelihood ratio statistic G^2 and the classical Pearson chi-square statistic X^2 can be used. Here, we use only the G^2 statistic.

The computational procedures of G^2 are given below.

(a) Compute

$$G^2 = 2\sum_{ijk} X_{ijk} \log\left(\frac{X_{ijk}}{\hat{m}_{ijk}}\right) ,$$

where $\{x_{ijk}\}$ are the observed counts, and $\{m_{ijk}\}$ are the fitted values obtained from the FIT table in 1.B.4, AS-51 algorithm.

(b) Look up the appropriate degree of freedom under the model assumption in Table 2.4. For example, under Model (a), the degree of freedom is (I-1)(J-1)(K-1).

(c) G^2 is distributed as a chi-square distribution with degrees of freedom appropriate for the model selected: For the G^2 value computed from (a), look for the probability of a value greater than the computed G^2 value in the chi-square table, Table A-6 in Appendix A, with the degrees of freedom obtained from (b).

1.B.6 A strategy for testing hypotheses.

The appropriate analysis of a multicontingency table usually involves several runs of the AS-51 algorithm, described in 1.B.4, unless one is interested in testing some particular hypothesis. Here, we perform two runs, as follows:

(a) First Run - Model Screening

The purpose of this run is to suggest interesting hypotheses for further analysis. The details of the procedure are given below.

PROCEDURE

(1) Compute the estimated cell counts $\{\hat{m}_{ijk}^{(1)}\}$ under model(a), using the AS-51 algorithm described in 1.B.4. Compute the G^2 statistic and look for its chi-square probability value according to 1.B.5. Record the values obtained.

(2) Compute the estimated cell counts $\{\hat{m}_{ijk}^{(2)}\}$ under Models (b.1), (b.2), and (b.3), using the AS-51 algorithm described in 1.B.4. For each table obtained, compute the corresponding

$$G^2(2/1)=-2\sum_{ijk} \hat{m}_{ijk}^{(1)}\log\frac{\hat{m}_{ijk}^{(2)}}{\hat{m}_{ijk}^{(1)}},$$

EXAMPLE

(1) Suppose, from the emission data of 1975-76 model-year cars in Element I, the G^2 statistic and its chi-square probability value under Model (a) are obtained according to procedures in 1.B.4 and 1.B.5 and recorded as follows:

Effect	D.F.	L.R.CHISQ	PROB
U_{123}	28	9.01	0.7020

(2) Under model (b.1) ($U_{13}=0$), computing the cell estimates and $G^2(2/1)$ gives $G^2(2/1)=7.79$. From the chi-square table (Appendix A, Table A-6) with degrees of freedom (2-1)(8-1)=7, get the probability value 0.0204. Similarly,

where $\{m_{ijk}^{(1)}\}$ is the contingency table obtained in (1). For each $G^2(2/1)$ obtained, look for its chi-square probability value with appropriate degrees of freedom. Under Model (b.1), this degree of freedom is (I-1)(K-1); under Model (b.2), it is (I-1)(J-1); under Model (b.3), it is (J-1)(K-1). Record, for each table computed, $G^2(2/1)$ degrees of freedom and the chi-square probability, including the table for U_{123} obtained in (1).

(3) Those effects whose probability values are less than 0.05 are considered to be significant (a value of 0.0000 indicates a highly significant effect), while those with probability value greater than 0.05 are considered to be nonsignificant and thus can be ignored for further analysis.

for Models (b.1) and (b.3), the $G^2(2/1)$'s and their chi-square value with appropriate degrees of freedom are obtained and recorded below (with the absent effect representing the model).

Effect	D.F.	L.R.CHISQ	PROB
U_{13}	7	7.79	0.0204
U_{12}	4	66.81	0.0000
U_{23}	28	27.23	0.1200
U_{123}	28	9.01	0.7020

(3) From the above summary table, one sees that the third-order interaction and the two-factor effect U_{23}(MAKE-CID interaction) are not significant and therefore are not needed in the final analysis. The two-factor interaction term U_{12} is highly significant, and thus belongs in the model. U_{13} is moderately significant, and so should probably be included in the model.

(b) Second Run - Testing Specified Models

Using the guidelines provided by the results of the first run, several models can be selected to fit the data. The models selected for testing include the highly significant terms found in the first run. These models follow the "hierarchy principle" that if any U-term is set equal to zero, all its high-order relatives should also be set equal to zero.

<table>
<tr><th>PROCEDURE</th><th>EXAMPLE</th></tr>
<tr><td>(1) Select several models which include the highly significant U-term found in the first run, (a), and which follow the hierarchy principle.</td><td>(1) Select Models $[U_{12}]$ and $[U_{12}, U_{13}]$, for testing, where the terms inside the brackets represent the highest-order interaction terms present in the model.</td></tr>
<tr><td>(2) For the selection models, estimate the expected cell counts according to procedures in 1.B.4 and compute G^2 and its chi-square statistics according to procedures in 1.B.5. Summarize the values into a table, with the top row representing the model with fewest U-terms and the second row representing the model with the next fewest U-terms, etc.</td><td>(2) Apply procedures in 1.B.4 and 1.B.5 separately to each of the selection models in (1) and produce the following table:
<table>
<tr><th>Model</th><th>G^2</th><th>Degrees of Freedom</th><th>Prob.</th></tr>
<tr><td>$[U_{12}]$</td><td>76.71</td><td>63</td><td>0.9</td></tr>
<tr><td>$[U_{12}, U_{13}]$</td><td>55.83</td><td>56</td><td>0.29</td></tr>
</table></td></tr>
<tr><td>(3) For each model, check the significance of the fit by comparing the corresponding Prob value with 0.05. If Prob > 0.05, the fit of the model</td><td>(3) Comparing Prob with 0.05, one finds that both $[U_{12}]$ and $[U_{12}, U_{13}]$ are nonsignificant; therefore, both models are sufficient to explain the</td></tr>
</table>

is nonsignificant; therefore, the model is sufficient to explain the relationships between the factors. If Prob < 0.05, then the fit is significant, and the model is not accepted as a good fit.

(4) Suppose Model (a) and Model (b) are both nonsignificant as concluded from (3), and Model (a) has fewer U-terms than Model (b). A way to decide which one of the two should be included in the final model can be described as follows: Let G_a^2 and G_b^2 be the G^2 values of Model (a) and Model (b), respectively. Let d_a and d_b be the corresponding degrees of freedom. Compute

$$D = G_a^2 - G_b^2 \text{ and } d = d_a - d_b .$$

Then, from the chi-square table, Table A-6, find p, the percentile of D with d degrees of freedom. If D is significant, i.e., $p < 0.05$, then Model (b) is considered to be a better one, and

relationships between the factors.

(4) Both $[U_{12}]$ and $[U_{12}, U_{13}]$ are nonsignificant. The difference between the fits of Models $[U_{12}]$ and $[U_{12}, U_{13}]$ gives a test of the two-factor interaction U_{13}; the G^2 value for the test is

$$76.71 - 55.83 = 20.88,$$

with $63 - 56 = 7$ degrees of freedom, where 76.71 is the G^2 value of $[U_{12}]$, and 55.83 is the G^2 value of $[U_{12}, U_{13}]$. The corresponding chi-square percentile of 20.88 with 7 degrees of freedom is 0.025. Since this percentile is less than 0.05, the test is significant, and therefore $[U_{12}, U_{13}]$ is the final model. In other words, CID and MAKE do affect the classification result.

therefore shall be included in the final model. On the other hand, if D is nonsignificant, i.e., $p>0.05$, then use Model (a).

(5) The magnitude of these U-terms can be computed from the FIT table of the model selection according to the following formulas:

The grand mean is given by

$U = (\Sigma \ell_{ijk})/IJK$.

The main effects are given by

$$U_{1(i)} = \frac{\ell_{i++}}{JK} - U$$

$$U_{2(j)} = \frac{\ell_{+j+}}{IK} - U$$

$$U_{3(k)} = \frac{\ell_{++k}}{IJ} - U \quad ,$$

where $\ell_{i++} = \sum_{jk} \ell_{ijk}$, $\ell_{+j+} = \sum_{ik} \ell_{ijk}$ and $\ell_{++k} = \sum_{ij} \ell_{ijk}$.

The interaction terms are given by

$$U_{12(ij)} = \frac{\ell_{ij+}}{K} - \frac{\ell_{i++}}{JK} - \frac{\ell_{+j+}}{IK} + U$$

$$U_{13(ik)} = \frac{\ell_{i+k}}{J} - \frac{\ell_{i++}}{JK} - \frac{\ell_{++k}}{IJ} + U$$

$$U_{23(jk)} = \frac{\ell_{+jk}}{I} - \frac{\ell_{+j+}}{IK} - \frac{\ell_{++k}}{IJ} + U$$

(5) Compute the U-terms from the FIT table $[U_{12}, U_{13}]$, $\{\hat{m}_{ijk}\}$ according to the formulas given on the left. One obtains, for example,

$U_{12(11)} = 2.5$, $U_{12(12)} = -0.5$,...etc.

The interpretation of these U-terms can be illustrated as follows: From $U_{12(11)} = 2.5$, which is positive, one concludes that the first CID level has a positive interaction with the number of misclassified cars at "every" MAKE level. From $U_{12(12)} = -0.5$, which is negative, one concludes that the second CID level has a negative interaction with the number of misclassified cars at every MAKE level.

NOTE: $\{\ell_{ij+}\}$ is the table of sums C_{12}.
Similarly $\{\ell_{i+j}\}$ is C_{13}, $\{\ell_{+jk}\}$ is C_{23},
$\{\ell_{i++}\}$ is C_1, $\{\ell_{+j+}\}$ is C_2, and $\{\ell_{++k}\}$
is C_3.

In the foregoing analysis, we used raw data only; we did not take into account a sampling factor. Below is a description of an adjusted table to account for the sampling differences.

Let $X_{1jk\ell}$ and $X_{0jk\ell}$ be the number of misclassified and correctly classified cars, respectively, in the $jk\ell$ cells. Let $n_{jk\ell}=X_{1jk\ell}+X_{0jk\ell}$. Thus $X_{+jk}=\sum_{\ell=1}^{44} n_{jk\ell}$. From the MVD file, find the population sizes, N_ℓ, of the ℓ^{th} vehicle groups, where $\ell=1,\dots,44$. The adjusted table $\{y_{ijk}\}$ is obtained by computing

$$y_{ijk} = \sum_{\ell=1}^{44} \left(X_{ijk\ell}/n_{jk\ell}\right) \cdot N_\ell \quad ,$$

where y_{ijk} is the expected number of i^{th} classified cars in the j^{th} MAKE, k^{th} CID.

Using the adjusted table $\{y_{ijk}\}$, go from 1.B.1 through 1.B.6 for the testing hypothesis. In case the classification does depend on MAKE and/or CID, redesign the classification rule according to MAKE and/or CID levels. Compute the E^* from Stage 4 for each level and compare it with the E^* of the current design, thus deciding the relative merit of the rules.

2.1.5.2 Element II Fleet. The Element II vehicle fleet for the Portland Study consists of 660 cars: 220 vehicles (110 1972-74 and 110 1975-77) recruited from Eugene/Springfield; 440 recruited from Portland. Of the 440 cars from Portland, 220 were of the 1972-1974 group and 220 were of the 1975-1977 group. In each of these two groups, 110 cars passed, and 110 failed, the Portland idle inspection test. Vehicles failing the idle inspection test are required to be reinspected following maintenance. All Element II vehicles are required to be tested once every three months for one year.

For the 1972-74 vehicle group, the classification algorithm of Stages 3 and 4 can be employed. The hypothesis on whether the error rates depend on the passing or failing of the idle test can be tested according to the 1.B procedures described in Section 2.1.5.1.

For the 1975-77 vehicle group, the classification rule designed for the Element I fleet can be applied, and the resulting error rates can be tested using 1.B procedures to see whether the idle test or the three-month recall would affect the error rates.

For Eugene cars, the classification rule of the Element I fleet can be applied directly, and the resulting misclassified rates can be compared with those of the Element I fleet in order to judge the appropriateness of the extension of the rule to the Eugene area. The Eugene cars were recruited from a stratified sampling scheme based on the distribution of vehicles registered in the Portland area [Hamilton Test Systems 1977]; but the Element I fleet was recruited from a stratified sampling scheme based on emission technology. Therefore, the comparison mentioned above would make sense only if the misclassified rates are properly adjusted according to sampling factors. The procedure for adjusting rates, along with the comparison, is given below.

PROCEDURE

(1) Let the misclassified rate at the j^{th} line item be $r_j = x_{ij}/n_j$, where x_{ij} and n_j are the number of misclassified cars and the number of total sample cars in the j^{th} line item, respectively $(j=1,\ldots,J)$.
NOTE: If n_j is too small, there is a need to combine several line items into one group, based on technology and model-year similarities.

(2) From the State MVD file, find N_j, the total number of registrations in the j^{th} line item, $j=1,\ldots,J$, and N, the total number of registrations in the Eugene region.

(3) Given the sampling plan described in "Plan of Performance" [Hamilton Test Systems 1977], for each vehicle group of the Element I fleet, identify those line items which form the group. For example, in the i^{th} vehicle group, there are K line items, numbered by $j_1, \ldots j_k$; then the adjusted rate of Eugene vehicles with respect to the i^{th} vehicle group is

$$\hat{r}_i = r_{j1}\left(\frac{N_{ji}}{N}\right) + \ldots + r_{jk}\left(\frac{N_{jk}}{N}\right).$$

(4) For the Element I fleet, if r_i is the raw misclassified rate for the i^{th} vehicle group, the adjusted rate is

$$\bar{r}_i = r_i\left(\frac{M_i}{M}\right),$$

where M_i and M are the number of registrations in the i^{th} vehicle group and in the Portland study area, respectively.

(5) Let $\{\hat{r}_i\}$ and $\{\bar{r}_j\}$ be the two tables obtained from (3) amd (4). Compare the two tables by testing the hypothesis H_0: $\hat{r}_i=\bar{r}_i$ for each i, using the following test:

Let $\hat{s}_i = 1 - \hat{r}_i$ and $\bar{s}_i = 1 - \bar{r}_i$.

Compute

$$z = \frac{\hat{r}_i - \bar{r}_i}{\sqrt{\frac{\hat{r}_i\hat{s}_i}{N} + \frac{\bar{r}_i\bar{s}_i}{M}}}$$

From the normal table, Table A-3 in Appendix A, find the p value, ignoring the sign of z. If this value is greater than .025, reject the hypothesis, i.e., $\hat{r}_i \neq \bar{r}_i$; otherwise, accept the hypothesis.

If a marked difference is shown in pattern and magnitude, then the Element I rule has limitations for being extended to the Eugene region. Therefore, a separate rule for Eugene may be necessary. This difference may be due to confounding factors, such as meteorological, geographical, and population variables. On the other hand, if no difference is shown, then Portland and Eugene are similar as far as the classification rule is concerned.

3.0 RESPONSES TO SPECIFIC QUESTIONS

The major objective of the Portland Study is to investigate the following three problems:

- ST/FTP association
- Immediate effectiveness of an Inspection/Maintenance program
- Long-term effectiveness of an Inspection/Maintenance program.

Within each problem, a series of specific questions or issues is raised by the investigators. The purpose of this chapter is to resolve, or to at least clarify, these questions or issues by providing either computational procedures or comments.

3.1 RESPONSES TO SPECIFIC QUESTIONS RELATED TO THE ST/FTP ASSOCIATION

3.1.1 Determination of Cutpoints

Methods presented in Stage 3, Section 2.1.3, produce a cutpoint (cutoff point), or cutpoints, for the prediction of FTP categorization. Under normal assumptions, a linear rank score for the short-test (ST) emission vector is computed, and the classification is done by means of a single cutpoint for the linear score. Under a non-normality assumption, a tree-structured decision rule is constructed and the classification is done by means of a series of individual pollutant cutpoints. In general, determination of these cutpoints is optimal; that is, the determined cutpoints achieve the minimum expected probability of misclassification for a given data base.

Specific questions related to cutpoints are the following:

1. Can the nine measurements obtained from the Federal three-mode test be fully utilized to calculate cutoff points?
2. Can pollutant-specific cutpoints be determined?
3. How should one determine if cutoff points are dependent on various engine parameters?

The answers to these questions are given below.

3.1.1.1 Federal three-mode test, nine measurements. For the Federal three-mode tests, three emission measurements are taken per test for each pollutant. Since the sequential approach to analysis described in Chapter 2 is readily applicable to high-dimensional data, these nine measurements can be fully utilized in the analysis of the ST/FTP association. Thus, under the joint normality condition, the stepwise discriminant analysis algorithm in Stage 3 will produce a linear rank score involving, at most, nine variables; under the joint non-normality condition, Friedman's algorithm will produce a tree-structured decision rule based on nine variables. The Friedman algorithm, incidentally, has two advantages for classifying high-dimensional data: efficiency of computation and ease of interpretation. Therefore, the user should try to use the Friedman's algorithm regardless of the outcome of the tests for joint normality in Stage 2 (pp. 2-24 through 2-35).

3.1.1.2 Pollutant-specific cutoff points. Individual pollutant-specific cutpoints can be obtained from the classification cutpoints in a straightforward fashion. Let $X = (X_1,\ldots,X_t)$ be ST emission variables, where t varies, depending on the type of emission test (number of pollutants tested and number of modes used to test each pollutant).

A. When the stepwise discriminant analysis algorithm is implemented, a best subset, say, $\{X_1,\ldots,X_k\}$, is selected and the Fisher discriminant function based on these k variables is computed. Typically, the classification rule will have the form: "Classify a car with emission vector $X = (X_1,\ldots,X_t)$ into Category I if

$$a_1 X_1 + \ldots + a_k X_k \geq d_k,"$$

where a_i, $i=1,\ldots,k$, and d_k are obtained from the printout (for details of the discussion, see Section 2.1.3, pp. 2-40 through 2-50). The individual cutoff points for $X_1,\ldots,X_k$ can now be defined as $d_k/a_1,\ldots,d_k/a_k$, respectively. For the rest of the variables, X_i, $i=k+1,\ldots,t$, the determination of individual cutoff points is not important but, for the sake of completeness, can be obtained from the printout, as follows:

First, search for the classification function in which all the variables are included (p. 2-44); e.g., classify an $\overline{X}$ as belonging to Category I if

$$b_1X_1 + \ldots + b_tX_t \geq d_t \quad ,$$

then define individual cutoff points for X_{k+1} as $d_t/b_{k+1},\ldots,X_t$ as d_t/b_t.

The above individual pollutant cutoff points can be best utilized to screen test cars, i.e., a car should be assigned to the fail FTP category whenever one or more of the ST emission values, $X_1,\ldots,X_k$, exceeds the cutoff points. Only if all the $X_1,\ldots,X_k$ emission values are below the cutoff points will the car's linear rank score be computed and a score-based classification be made.

B. When the Friedman's algorithm is implemented, a series of pollutant-specific cutoff points is given explicitly in the printout (see Friedman's algorithm, Section 2.1.3).

3.1.1.3 Cutpoints and engine parameters. For the purpose of testing whether the cutoff points discussed above depend on various engine parameters, the user should consult Stage 5, Section 2.1.5, which deals with hypothesis testing.

3.1.2 Determination of Error Rates and 95% Confidence Interval

Questions related to error rates and 95% confidence interval are the following:

1. Provide a means of estimating the expected probability of misclassification and its 95% confidence interval.
2. How could a 95% confidence interval around the individual pollutant cutpoints be calculated?

The answers to these questions are given below.

3.1.2.1 Probability of misclassification and its 95% confidence interval. Stage 4 gives a complete account of the estimation of the probabilities of errors and their 95% confidence intervals.

3.1.2.2 Individual pollutant cutoff points mentioned in 3.1.1. The 95% confidence interval around the individual pollutant cutoff points should be derived from the distributions of such cutoff points. But it is not at all clear why the points computed from the classification function should have a reasonable distribution, much less a normal one. In fact, the central concept here is the 95% confidence interval for the error rate, which gives an estimate of the "reliability" of the method being used. With the following ad hoc method, one can roughly estimate the 95% confidence interval around the pollutant cutoff points:

Systematically leave out 10% of the data points and use the rest of the points to construct the classification scheme according to the methods in Stage 3, Section 2.1.3, and compute the cutoff point as in Section 3.1.1.1.

A. For the discriminant analysis algorithm, three cutoff points are produced:

TRIAL	CUTOFF POINTS		
1	d_1/a_1	d_1/b_1	d_1/c_1
.	.	.	.
.	.	.	.
.	.	.	.
k	d_k/a_k	d_k/b_k	d_k/c_k

where k denotes the number of trials. For the i^{th} column, i=1,2,3, compute the mean and standard deviation of the k numbers, say, μ_i, σ_i. Then the 95% confidence interval for each of the three cutoff points is given by

$$\mu_i \pm t_{.95}\,\sigma_i \qquad \text{for } i=1,2,3,$$

where $t_{.95}$ is the 95th percentile from the t distribution with k-1 degrees of freedom (see Table A-4, Appendix A).

B. The same method can be applied to Friedman's algorithm, but the results are meaningless.

3.1.3 Determination of Sample Sizes

Questions related to determination of sample sizes are the following:

1. What is the minimum number of vehicles/tests required for fixed probability of misclassification in the range of 5% to 20%.
2. What is the minimum number of vehicles/tests required for the estimates of probability of misclassification with 95% confidence.
3. Can a minimum number of vehicles/tests for the simultaneous achievement of 1 and 2 be determined?

The answers to these questions are given below.

3.1.3.1 For fixed probability of misclassification. The purpose of this subsection is to estimate the minimum number of vehicles/tests required for fixed probability of misclassification, E^*, in the range of 5% to 20%. Here, $E^* = pE_1 + (1-p)E_0$, where p is the probability that an observation comes from Category I; and E_1 and E_0 are the probabilities of false negative error and false positive error, respectively. The calculations are broken into two steps.

Step 1. Run a pilot study to estimate p and the sample Mahalanobis distance D^2 according to

Estimation of p

Suppose that, in the pilot study, n cars undergo a series of ST and FTP. For a given set of FTP standards, if n_1 cars pass the FTP and $n_2=n-n_1$ cars fail the FTP, estimate p by n_1/n.

Estimation of D^2

Now let $Y_i(i=1,\ldots,n_1)$ be the observations of ST emissions for cars passing the FTP and $Z_i(i=1,\ldots,n_2)$ be the observations of ST emissions for cars failing the FTP. Then compute

$$(1)\quad \bar{Y} = (1/n_1)\sum_{i=1}^{n_1} Y_i,\quad \bar{Z} = (1/n_2)\sum_{i=1}^{n_2} Z_i \text{ and}$$

$$S = \frac{1}{n-2}\left\{\sum_{i=1}^{n_1}(Y_i-\bar{Y})(Y_i-\bar{Y})^T + \sum_{i=1}^{n_2}(Z_i-\bar{Z})(Z_i-\bar{Z})^T\right\}$$

$$(2)\quad D^2 = (\bar{Y}-\bar{Z})^T S^{-1}(\bar{Y}-\bar{Z})$$

(A numerical example has been given in Section 2.1.3, pp 2-38 through 2-40.)

Step 2. Let N be the minimum number of vehicles to be sampled in the formal study. An estimate of N can be obtained through the following computational procedure, which is derived from a theoretical expression of the probability of misclassification given by Okamoto [1963]:

PROCEDURE	EXAMPLE
(1) Set e = fixed level of probability of misclassification.	(1) Set $e = 0.15$
(2) Estimate p and D^2 according to Step 1.	(2) Suppose, from the pilot study of Step 1, that $p=\frac{2}{3}$, $D^2=1.2$
(3) Compute $c = \left\{\log\left\lvert\frac{(1-p)}{p}\right\rvert - \frac{D^2}{2}\right\}/D$	(3) $c=(\log(.5) - 0.6)/1.09$ $=-1.186$
(4) Let k be the number of emission variables in the study. Compute	(4) Suppose there are three emission variables under consideration. Then
$a_1 = 3c-c^3+k(D-c)$	$a_1 = 4.938$
$a_2 = 2D+3c-c(D+c)^2-k(D+c)$	$a_2 = -0.297$
$a_3 = 2(D+3c)-c(D+2c)^2-2k(D+3c)$	$a_3 = 8.351$

(5) Let Φ and ϕ denote the cumulative distribution function and the density function, respectively, of the standard normal distribution. Look up $\Phi(c)$ from the normal table (Table A-3); compute $\phi(c) = \frac{1}{\sqrt{2\pi}} e^{-c^2/2}$	(5) Look up $\Phi(-1.186)$ from the normal table and get $\Phi(-1.186) = 1 - \Phi(1.186)$ $= 1 - 0.8810$ $= 0.1190$ $\phi(-1.186) = 0.2792$
(6) Set $p_1(c,D)=\Phi(c)+(\phi(c)/N)\left(\frac{a_1}{2pD^2} + \frac{a_2}{2(1-p)D^2} + \frac{a_3}{4}\right)$ $p_0(c,D)=\Phi(c)-(\phi(c)/N)\left(\frac{a_2}{2pD^2} + \frac{a_1}{2(1-p)D^2} + \frac{a_3}{4}\right)$	(6) Set $p_1 = 0.1190 + (13.81)/N$ $P_0 = 0.1190 - (6.647)/N$
(7) Set the equation $e=pP_1(c,D) + (1-p)P_0(c,D)$ and solve for N, where e is given in (1).	(7) Set $0.15 = 0.1190 + (6.99)/N$ Therefore, $N \approx 225$, which is the minimum number of cars required, given the level of probability of misclassification as 15%.

CAUTION: The precision of this estimate depends on 1) the assumption that the emission variables behave like a joint normal distribution and 2) the assumption that the estimated D^2 and p in Step 1 are close to their true values.

3.1.3.2 For a given length of 95% confidence interval. The purpose of this section is to estimate the minimum number of vehicles/tests required for the estimates of probability of misclassification to fall within ± X% of the true probability of misclassification with 95% confidence. The procedure for obtaining this minimum sample size is based on, first of all, a method developed by McLachlan [1975] for constructing an accurate confidence interval for the individual probabilities of errors, E_1 and E_0, and second, the Bonferoni inequality, which relates the two individual confidence intervals to a joint confidence region of the two probabilities of errors (discussed in Stage 4, Section 2.1.4, p. 2-68). The details of the computational procedures are described below (the notations and the example used in the previous subsection are carried over here).

	PROCEDURE		EXAMPLE
(1)	Set a fixed X% level, and a fixed $100(1-\alpha)\%$ level of confidence interval.	(1)	Set X% = 5%, $\alpha=.05$
(2)	Copy the values of p, D^2 estimated from Step 1 and c from Step 2(3), in Section 3.1.3.1.	(2)	From the pilot study, it has been estimated that $p=\frac{2}{3}$, $D^2=1.2$. Thus, $c=-1.186$ from Step 2(3).
(3)	Compute $b_1=D^2+4(3k-4)+(k^2-4k+5)(16/D^2)$ $b_2=(D^2-2k)/8$ $b_3=D^4+2(11k-16)D^2+8(5k-4)$ $b_4=D^4-2(k+4)D^2+8k$ $b_5=2D^6+16(2k-5)D^4-32(4k-13)D^2$	(3)	$k=3$; therefore, $b_1 = 47.86$ $b_2 = 0.6$ $b_3 = 130.8736$ $b_4 = 9.2736$ $b_5 = 77.5495$

PROCEDURE	EXAMPLE
(4) Compute $\phi(c) = \frac{1}{\sqrt{2\pi}} e^{-c^2/2}$	(4) $\phi(-1.186) = 0.2792$
(5) Set $\sigma_1^2(N)=(\phi(c))^2[(\frac{1}{p}+\frac{D^2}{8})/N+b_1/(4pN)^2$ $+b_2/(p(1-p)N^2)+b_3/(64pN^2)$ $+b_4/(64(1-p)N^2)+b_5/(32N)^2]$	(5) Set $\sigma_1^2(N) = \frac{0.037}{N} + \frac{1.260}{N^2}$
$\sigma_2^2(N)=(\phi(c))^2[(\frac{1}{1-p} + \frac{D^2}{8})/N+b_1/(4(1-p)N)^2$ $+b_2/(p(1-p)N^2) + b_3/(64(1-p)N^2)$ $+b_4/(64pN^2)+b_5/(32N)^2]$.	$\sigma_2^2(N) = \frac{2.425}{N} + \frac{2.77}{N^2}$
(6) Let Z_α be the 100α percentile of the standard normal distribution (can be found in Table A-3, Appendix A). Find the minimum N which satisfies the inequalities $\sigma_1(N)\ Z_{\alpha/4} \leq X\%$ and $\sigma_2(N)\ Z_{\alpha/4} \leq X\%$ simultaneously; this N will be the minimum number of vehicles required for the estimates of probability of misclassification to fall within ±5% of the true probability of misclassification.	(6) $\alpha/4 = 0.05/4 = 0.012$ From Table A-3, $Z_{0.012} = 0.03$ Find the minimum number N which satisfies the inequalities $\frac{0.0001}{N} + \frac{0.037}{N^2} \leq 0.05$ $\frac{0.072}{N} + \frac{0.002}{N^2} \leq 0.05$ simultaneously. Equivalently, the number N can be obtained by, first, finding the solutions of the equations $(0.001)N+0.037-(0.05)N^2=0$ and $(0.072)N+0.002-(0.05)N^2=0$;

PROCEDURE

EXAMPLE

and, second, by comparing the two sets of solutions. The maximum number obtained is the answer. The solutions of the above two equations are $N_1=2.7$ and $N_2=8.2$, respectively; therefore N=9 is the minimum number of vehicles required for the estimates of probability of misclassification to fall within ± 5% of the true probability of misclassification.

COMMENT: For individual cutoff points as defined in 3.1.1, the minimum sample sizes required for an accurate estimation cannot be determined, mainly because the distributional properties for such cutoff points are unknown or are too complex.

3.1.3.3 Combination of Sections 3.1.3.1 and 3.1.3.2. A minimum number of vehicles/tests for the simultaneous achievement of Sections 3.1.3.1 and 3.1.3.2 can be determined by simply taking the maximum of the two individual numbers obtained in those sections.

3.1.4

Develop a method for comparing the ST(laboratory)/FTP association with the ST(real-world)/FTP association, considering possible places for human/instrument error in both situations.

Assume that the ST(laboratory)/FTP data is available. Then perform a classification according to Stage 3 (Section 2.1.3) and compute the expected probability of misclassification, E_1^*, according to Stage 4 (Section 2.1.4). Do the same for ST(real-world)/FTP data and obtain the expected probability of misclassification, E_2^*. Compare these two numbers. The lower the number, the better the association.

3.2 RESPONSES TO QUESTIONS RELATED TO IMMEDIATE EFFECTIVENESS OF AN INSPECTION/MAINTENANCE PROGRAM

Background

The Element II fleet in the Portland Study includes 660 vehicles of model years 1972-1977, 440 of which are registered in the Portland area (and thus are subject to inspection/maintenance, I/M), and 220 of which are operated in Eugene, Oregon, where no I/M program exists. The set of 440 vehicles consists of 110 1975-1977 model-year cars which failed the Oregon state inspection test and 110 which passed; and 110 1972-1974 model-year cars which passed and 110 which failed. The 220 Eugene cars were selected to be representative of the Portland population of registered 1972-1977 cars.

Each of the 660 cars undergoes an initial series of short tests at a Portland State inspection station (the same station is used for all vehicles). Following this series of tests, each vehicle undergoes another set of short tests, plus additional emission and fuel economy tests at the contractor's emission testing laboratory. Each vehicle that failed Portland's idle emission test at the State facility undergoes maintenance and is retested on the idle test at the State inspection station; and is retested over a series of emission/fuel economy tests at the contractor's laboratory.

3.2.1

Suggest ways to determine before-I/M emission levels and confidence limits for the Portland area vehicles, for each model-year vehicle from 1972-1974, using the sample of data available; also determine these values for the Eugene area and/or the parts of Oregon not requiring I/M.

Assume that the before-I/M emission levels for the Portland area 1972-1974 cars are not available directly from some other source. Then levels can be estimated from the Eugene emission data for 1972-1974 cars, using the weighted average technique described below.

PROCEDURE

(1) Choose a specific pollutant emission variable, say, X.

(2) Suppose there are K MAKE-YEAR combination groups [Hamilton Test Systems 1977], e.g., DODGE-72. From the DMV registration file, find the total number of cars in the i^{th} MAKE-YEAR group, say, N_i, $i=1,\ldots,K$. Let N be the total number of 1972-74 cars in Eugene.

(3) For the sample of Eugene cars in the i^{th} MAKE-YEAR group, compute the mean and variance of the variable X, say, Y_i and S_i^2, $i=1,\ldots,K$, respectively.

(4) Compute the weighted average

$$\overline{Y}_X = \sum_{i=1}^{K} Y_i\left(\frac{N_i}{N}\right)$$

and variance

$$S_X^2 = \sum_{i=1}^{K} S_i^2\left(\frac{N_i}{N}\right)^2 .$$

(5) Use $\overline{Y}_X$ as an estimate of the before-I/M emission levels for the Portland area for each model-year vehicle from 1972-1974. The confidence limit would be set as

$$\overline{Y}_X \pm S_X \cdot t_{.95}(n-1) \quad ,$$

where $t_{.95}$ is the 95th percentile from the t distribution with n-1 degrees of freedom; and n is the total number of sample points. For $n > 120$, use $t_{.95} = 1.96$ (which follows from the normal approximation).

The validity of using $\overline{Y}_X$ to estimate the before-I/M X emission levels for the Portland area can be further checked by comparing the value with the national emission value, if the latter is available. Thus, the national emission level's falling within the confidence limits of $\overline{Y}_X$ means that the extrapolation from Eugene to Portland is valid. Otherwise, the extrapolation is less justifiable and one should be conservative in making the conclusion.

3.2.2

3.2.2.1 Discuss ways to determine if the State inspection test failure rates for similar groups of vehicles are different in the statistical sense. Do failure rates differ from Eugene to Portland? See Section 2.1.5, in which hypotheses about the failure rates are tested.

3.2.2.2 Even if the failure rates for Eugene and Portland are similar, it is possible that the emission levels of failed/passed vehicles are different. Suggest ways to look at this possibility. Analysis of variance can be used. For details of the procedure, see Section 3.2.7, pp. 3-22 and 3-23.

3.2.3

3.2.3.1 Suggest methods to determine if the change in emission levels and fuel economy following maintenance is significant for the Portland failed cars and for the Eugene failed cars. The standard way of testing whether the mean is equal to zero will be used. The details of the procedure are given below.

PROCEDURE	EXAMPLE
(1) Select a variable of interest.	(1) For the Portland failed cars, select the short-test NO_x variable.

PROCEDURE	EXAMPLE
(2) For the i^{th} car, compute $Y_i=X_i^{(a)}-X_i^{(b)}$, where $X_i^{(a)}$ and $X_i^{(b)}$ are the emission levels after and before maintenance, respectively.	(2) For each Portland failed car, compute $Y_i=X_i^{(a)}-X_i^{(b)}$, where $X_i^{(a)}$ and $X_i^{(b)}$ are the ST-NO_x values after and before maintenance, respectively.
(3) Use the SPSS-CONDESCRIPTIVE or BMDP1D subroutine to compute the sample mean, $\hat{\mu}$, and the sample standard deviation, $\hat{\sigma}$, for the Y_i's.	(3) Use the SPSS-CONDESCRIPTIVE subroutine for the Y_i's. From the printout, one can see that the mean is 2.3 and the standard deviation is 0.4.
(4) Test $H_0:\mu=0$. Find the 95% confidence interval for μ, which has the form $\hat{\mu} - 1.96\hat{\sigma} < x < \hat{\mu} + 1.96\hat{\sigma}$. The significance of the mean, μ, can be tested by examining the confidence interval to see whether the zero value is contained in the interval. If the zero is not contained in the interval, then μ is significant; otherwise, μ is nonsignificant.	(4) Test $H_0:\mu=0$ The 95% confidence interval for μ would be $2.3 - 1.96(.4) < x < 2.3 + 1.96(.4)$. Since 0 is not contained in the interval, one concludes that μ is significant; i.e., the change in NO_x levels following maintenance is significant for the Portland failed cars.

3.2.3.2 Can change in FTP be expressed as a function in ST? Let Y denote the change in FTP and X denote the change in ST. Perform a regression analysis according to the procedure in Section 3.2.5, with Y as dependent variable and X as independent variable.

3.2.3.3 Do highway, FSC, HFET, and FTP fuel economy all change significantly? For any of the selected variables, let $Y = Y^{(b)} - Y^{(a)}$, where $Y^{(b)}$ and $Y^{(a)}$ are the values obtained before and after the maintenance, respectively. Perform the test for the significance of the mean according to the procedure described in Section 3.2.3.1. If the test is significant, then one can conclude that the change is significant.

3.2.3.4 Suggest a method for putting confidence limits around the before-I/M levels, the after-I/M levels, and the change in levels. Let Y be the level of interest, either the before- or after-I/M level, or the change in levels. Using the descriptive subroutine (Section 2.1.1, Stage 1), compute the mean, $\hat{\mu}$, and standard deviation, $\hat{\sigma}$, from the data. Then the 95% confidence limit around the level of interest is given by

$$\hat{\mu} - \hat{\sigma} t_{97.5}(\nu) \leq x \leq \hat{\mu} + \hat{\sigma} t_{97.5}(\nu) \quad ,$$

where $t_{97.5}(\nu)$ is the 97.5th percentile of the t distribution (Table A-4, Appendix A) with degrees of freedom [ν = (number of sample points) - 2].

3.2.3.5 Discuss ways to determine whether cost of maintenance is dependent on a vehicle's amount, or percentage, above the short-test cutpoints. This problem can be solved by using the linear regression subroutine mentioned in Section 3.2.5. In this case, the dependent variable is cost of maintenance and the independent variable is the amount or percentage above the short-test cutpoint(s). The goodness-of-fit of the regressive line shall be judged by the magnitude of R^2. If the value of R^2 is greater than 0.6, the fit is considered significant: The higher the R^2, the better the fit.

3.2.4

Suggest ways to determine if maintenance performed because of ST failure on one pollutant affects ST and FTP emission levels on other pollutants.

The approach to this problem depends on the type of maintenance performed and the ST cutoff points.

Assume that the maintenance protocol is well established, and the cutoff point for short-test emission value on one pollutant, for example, HC, is available from Section 3.1.1; then, for cars which fail the short test, perform the following steps:

1. Compute the difference in idle CO level before and after maintenance,

$$Y_{CO} = Y_{CO}^{(a)} - Y_{CO}^{(b)} \quad .$$

2. Perform the test for the mean of Y_{CO} equal to zero according to procedures described in Section 3.2.3.1.

The same method can be applied for other pollutants.

3.2.5

Discuss ways to determine whether cost of maintenance is dependent on a vehicle's amount or percentage above the short-test cutpoint(s).

A linear regression analysis can be applied to the problem. The details of the analysis are given below.

MULTIPLE LINEAR REGRESSION

(1) Select variables of interest, one as the dependent variable and the rest as independent variables. Let Y be the dependent variable and $X_1,\ldots,X_p$ be the independent variables. Then the multiple linear regression model is

$$Y = b_0 + b_1X_1 + \ldots +b_pX_p \quad ,$$

where $b_0,b_1,\ldots,b_p$ are the regression coefficients to be determined.

<u>EXAMPLE</u>: Select cost of maintenance, C, as the dependent variable and select a vehicle's amounts, X_1, X_2 and X_3, above the three pollutant-specific short-test cutpoints as independent variables. The model is $C = b_0 + b_1X_1 + b_2X_2 + b_3X_3$.

(2) Perform a multiple linear regression analysis through the use of BMDP1R or SPSS MULTIPLE REGRESSION ANALYSIS: SUBPROGRAM REGRESSION subroutines, with C as the dependent variable and X_1, X_2 and X_3 as independent variables. A typical output example, along with its interpretation, is given below

OUTPUT EXAMPLE FOR MULTIPLE LINEAR REGRESSION (BMDP1R)

```
                              .
                              .
                              .

     MULTIPLE R           0.9741      STD. ERROR OF EST.   2.5692

     MULTIPLE R-SQUARE    0.9488

ANALYSIS OF VARIANCE

               SUM OF SQUARES    DF    MEAN SQUARE   F RATIO    P(TAIL)

REGRESSION           X            X         X         98.822     0.0000

RESIDUAL             X            X         X
```

SUMMARY TABLE:

VARIABLE	COEFFICIENT	STD ERROR	STD. REG. COEFFICIENT	T	P(2 TAIL)
Constant	-43.7040				
X1	0.889	0.119	0.779	7.481	0.000
X2	0.817	0.325	0.253	2.512	0.023
X3	-0.107	0.125	-0.055	-0.860	0.402

INTERPRETATION OF THE OUTPUT EXAMPLE

PROCEDURE

(1) From the printout, find the regression coefficients summary table, which includes regression coefficient, b_i, standard error, $Se(b_i)$, t-statistics (in SPSS, this is F-statistics) and the p value for the i^{th} variable X_i, where $i=1,\ldots,p$.

(2) The linear regression equation is given by

$$Y = b_0 + b_1X_1 + \ldots + b_pX_p$$

where b_i's are the regression coefficients found in (1).

EXAMPLE

(1) From the OUTPUT EXAMPLE, one finds that the regression coefficients summary table is as follows:

VARIABLE	COEFFICIENT	STD. ERROR	p
CONSTANT	-43.7040		
X_1	0.889	0.119	0.000
X_2	0.817	0.325	0.023
X_3	-0.107	0.125	0.402

(2) The linear regression equation is given by

$$Y = -43.7040 + 0.889(X_1) + 0.817(X_2) - 0.107(X_3)$$

PROCEDURE	EXAMPLE
(3) For each coefficient b_i, test the hypothesis $H_0: b_i = 0$. The acceptance of this hypothesis can be interpreted as "the variable X_i does not significantly improve the prediction of Y over that obtained by regressing Y on the other p-1 variables." The test results can be obtained immediately by comparing the p value found in the last column of the table in (1) with the value 0.05. If the p value is greater than 0.05, the hypothesis that $b_i = 0$ is accepted and thus the coefficient is significant; otherwise, the coefficient is significant at the 0.05 level. In SPSS-MULTIPLE REGRESSION ANALYSIS, the p value is not available. From the F-statistic, the p value is the area to the	(3) By observing the p values shown in the task in (1), one concludes that b_1 is highly significant, b_2 is moderately significant, and b_3 is insignificant. The interpretation is that, in the prediction of the cost of maintenance, X_1 contributes significantly to the equation. X_2 contributes moderately, but X_3 does not contribute to the equation in a significant way.

PROCEDURE	EXAMPLE
right of F under the $F(1,\nu_R)$ frequency curve (see Table A-5), where ν_R = n-p-1 degrees of freedom.	
(4) Look for the MULTIPLE R-SQUARE value for goodness-of-fit. A large multiple R^2 (greater than 0.6) indicates that the fit is good. The larger the R^2, the better the fit.	(4) $R^2 = 0.9488$. Therefore, one concludes that the linear fit is quite good.

3.2.6

Using the Eugene and Portland data, determine a way to analyze the effect of an I/M program's existence on the extent to which voluntary owner maintenance will occur.

In the Portland Study, car owners responded to questionnaires concerning maintenance. A scoring system must be designed for the questionnaire, so that a score for maintenance is associated with each vehicle. Once all scores are figured, the average score for vehicles in Portland and the average score for vehicles in Eugene can be computed. These two means (averages) can then be compared.

3.2.7

Suggest ways for determining whether the change in fuel economy/emissions following maintenance is dependent on vehicle characteristics such as CID, inertia weight, cutpoints to which the vehicle was subjected, engine family, and manufacturer.

When all the categorical variables are considered in the analysis simultaneously, classical analysis of variance is not suitable simply because it is not designed for effectively extracting information from large high-dimensional data sets. The situation, however, can be easily tackled by the Automatic Interaction Detection (AID) algorithm. AID is a generalized data analysis program which uses analysis-of-variance techniques to explain as much of the variance of a given dependent variable as possible. In fact, it makes successive dichotomous partitions on the sample, using independent variables to "predict" the dependent variable in such a way as to maximize differences among the split groups. Thus, a tree-structured output will be produced.

The MIDAS and TSC computing library contain an AID subroutine. For a complete program card setup and output examples, the user should read Sonquist, et al. [1975], "Searching for Structure," Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

3.2.8

Determine estimates and confidence limits for before-maintenance emission levels (FTP and ST) and fuel economy separately for passed and failed Portland vehicles by model-year group (1972-74 and 1975-77), and for passed and failed Eugene vehicles by the same model-year groupings. Determine similar estimates for failed cars after maintenance.

PROCEDURE

(1) From each sample-car owner, obtain information on the time of the last maintenance, along with the type of maintenance performed before entering the study. Let T be the difference between the time of entering the study and the time of last maintenance before the study.

(2) Select a variable of interest, say, X, and a group of cars in the study; e.g., ST-NO_x for passed Portland vehicles, by model-year 1972-74 group.

(3) From those cars with almost the same T (as defined in (1)), compute the mean, $\hat{\mu}$, and the standard deviation, $\hat{\sigma}$, of the X variable. The confidence limit is given by $\hat{\mu} \pm t_{.95}(\nu)\,\hat{\sigma}$, where $t_{.95}(\nu)$ equals the 95th percentile of t distribution with degrees of freedom [ν = (number of sample points)-1]. Thus, for various before-maintenance time levels, the estimates and confidence limits of emission levels are obtained.

Since each vehicle that failed Portland's idle emission test undergoes maintenance and is retested over a series of STs and FTPs, the estimates and confidence limits for failed cars after maintenance can be computed directly from the data.

3.2.9

Does the initial effectiveness of I/M on emissions and fuel economy depend on seasonal variations in temperature?

The solutions to this problem rely on the results of 3.2.1 and the data gathered from Element II fleet vehicles which were recruited from the Portland region (220 cars in the 1972-74 model-year group). The procedure is described below.

PROCEDURE

(1) Choose a variable of interest, say, X.

(2) Notice that the Element II cars are recalled every three months for one whole year. Let X_i^b and X_i^a be the average before-I/M and after-I/M levels at time i, respectively, where i=1,2,3, and 4. Let S_i^b and S_i^a be the corresponding standard deviation of the estimates.

(3) Let T_i be the average temperature at time i.

(4) Compute $r_i = X_i^a - X_i^b$, the effect of I/M on emissions.

(5) Plot r_i versus T_i, i=1,2,3, and 4, and look for a trend.

(6) Let N_b and N_a be the number of sample points used to determine X_i^b and X_i^a, respectively. Perform the two-sample Welch t test for $H_0{:}r_i=0$. The test statistics are

$$T_0 = r_i \Big/ \sqrt{\frac{(S_i^a)^2}{N_a} + \frac{(S_i^b)^2}{N_b}} \quad .$$

Find $t_{.95}(\nu)$, the 95th percentile of t distribution with ν degrees of freedom, where

$$\upsilon = \left(\frac{(s_i^a)^2}{N_a} + \frac{(s_i^b)^2}{N_b}\right)^2 \Bigg/ \left(\frac{(s_i^a)^4}{N_a^2(N_a-1)} + \frac{(s_i^b)^4}{N_b^2(N_b-1)}\right).$$

If $t_{.95}(\nu) < T_0$ (Table A-4), reject the hypothesis that $r_i=0$. (If $\nu \geq 120$, use 1.96 instead of $t_{.95}$.) Otherwise, accept it. Here, reject that $r_i=0$ means there is a significant I/M effect.

3.2.10

3.2.10.1 Does the probability of failing the idle test change with queue time?

The following procedures are recommended:

PROCEDURE

(1) Subdivide the queue time period into a finite number of equal subintervals, say, $I_1,\ldots,I_t$.

(2) The probability of failing the idle test when queue time falls into the interval I_i can be estimated by x_i/n_i, where x_i is the number of cars which fail the idle test and which have queue times falling in I_i, and n_i is the total number of cars with queue times falling in I_i.

(3) Do a regression analysis with p_i as the dependent variable, and test for slope = 0. (See Section 3.2.5 for details of the procedure.)

3.2.10.2 Does correlation between the idle test and FTP depend on queue time?

Owing to the complexity of the distribution for correlation, the easiest way is to plot the correlation according to the following procedure:

PROCEDURE

(1) Subdivide the queue time period into a finite number of equal subintervals, say, $I_i,\ldots,I_t$.

(2) Let r_i be the correlation between the idle test and the pollutant-specific FTP for cars with queue times falling in the interval I_i.

(3) Plot r_i versus the midpoint p_i of I_i.

(4) Do a regression analysis with r_i as dependent variable and p_i as independent variable, and test for slope = 0. (See Section 3.2.5 for details of the procedure.)

3.2.11

What percentage of the reduction in FTP emissions could have been achieved by a diagnositc inspection as opposed to an emission test?

Vehicle engine diagnostics are required for vehicles in the Element II program. There are twelve specific measurements to be made for diagnostic checks [Hamilton Test Systems 1977]. Therefore, apply Friedman's classification algorithm as described in Stage 3 and evaluate the classification result by computing the probability of misclassification according to Stage 4. Finally, compare this result with the ST/FTP classification result based on emission tests to get the percentage of the reduction in FTP emission that could have been achieved by a diagnostic inspection.

3.2.12

3.2.12.1 Discuss methods of evaluating the ST/FTP association with respect to fuel economy. The determination of fuel economy is based on measured CO_2. Thus, for each CO_2 level, say, level i, record the number of misclassified cars, n_i, and the total number of cars, m_i, based on a given classification scheme. Define the misclassified rate at level i as $p_i = n_i/m_i$. Do a regression analysis with p_i as the dependent variable and level i as the independent variable, and test for slope = 0. (See Section 3.2.5 for details of the procedure.)

3.2.12.2 Suggest methods for determining statistically significant relationship between pollutant-specific emission levels and fuel economy. Do a regression analysis with pollutant-specific emission level as the dependent variable and CO_2 as the independent variable, and test for slope = 0. (See Section 3.2.5 for details of the procedure.)

3.3 RESPONSES TO QUESTIONS RELATED TO LONG-TERM EFFECTIVENESS OF INSPECTION/MAINTENANCE

Background

Each of the 660 vehicles in Element II fleet will be returned to the contractor's laboratory facility 4 times, at 3-month intervals, and will be retested on the State inspection station test, the FTP, the hot-start FTP, and the fuel economy test.

3.3.1

3.3.1.1 Estimate short-test, FTP, and fuel economy deterioration over time, for Portland and Eugene cars. Suggest overall analysis for each model-year grouping (1972-74, and 1975-77), and also for passed vehicles and failed vehicles separately.

The following method can be used for estimating deterioration over time:

PROCEDURE	EXAMPLE
(1) Select a variable of interest, say, X, for a particular vehicle group.	(1) Select the short-test NO_x variable for Portland cars.
(2) Do a linear regression analysis with X as the dependent variable and time T_i, where i=1,2,3,4, as the independent variable. (The details of the procedure are given in Section 3.2.5). Record the regression coefficient, b, and	(2) Do a linear regression analysis with ST-NO_x as the dependent variable and time, T_i where i=1,2,3 and 4, as the independent variable. From the printout (see Section 3.2.5), one finds that b=2.3

its standard error, Se(b), from the output. The interpretation of b is "the change per unit time."	and Se(b) = 0.4. Thus, the change per unit time of the ST-NO$_x$ is 2.3.
(3) Compute the 95% confidence interval according to the formula $b-Se(b)\cdot t_{97.5}(\nu)\leq x\leq b+Se(b)\cdot t_{97.5}(\nu)$, where $t_{97.5}(\nu)$ can be found from the percentile of the t distribution table with degrees of freedom $\nu=n-2$, (n is the number of sample points). For n larger than 120, use $t_{97.5} = 1.96$.	(3) The 95% confidence interval is $2.3-1.96(.4)<x<2.3+1.96(.4)$. (The number of sample points is larger than 120). Since the interval does not contain zero, the coefficient b is significant at 0.05 level.

3.3.1.2 Is there a significant difference between pollutant-specific Portland and Eugene slopes? Is there a significant difference between slopes for passed versus failed vehicles? Two methods can be used to solve the problem.

Method 1: Let the observations of a specific pollutant for N cars in Portland at time i be $X_1^{(i)},\ldots,X_N^{(i)}$, distributed according to a normal distribution with mean $\mu_x^{(i)}$ and standard deviation $\sigma_x^{(i)}$, where i=1,2,3, and 4. Similarly, let the observation of the specific pollutant for M cars in Eugene at time i be $Y_1^{(i)},\ldots,Y_M^{(i)}$, distributed according to a normal distribution with mean $\mu_y^{(i)}$ and standard deviation $\sigma_y^{(i)}$, i=1,2,3, and 4. The problem of testing the difference between Portland and Eugene emission slopes reduces to testing the following hypotheses:

$$H_0^{(1)}: \mu_x^{(1)} - \mu_x^{(2)} = \mu_y^{(1)} - \mu_y^{(2)}$$

$$H_0^{(2)}: \mu_x^{(2)} - \mu_x^{(3)} = \mu_y^{(2)} - \mu_y^{(3)}$$

$$H_0^{(3)}: \mu_x^{(3)} - \mu_x^{(4)} = \mu_y^{(3)} - \mu_y^{(4)}$$

Using the test for $H_0^{(1)}$ as an illustration, the details of the procedure are given below.

PROCEDURE	EXAMPLE
(1) For each car j, $j=1,\ldots,N$, in Portland, compute $U_j = X_j^{(1)} - X_j^{(2)}$. Similarly, for each car K, $K=1,\ldots,M$, in Eugene, compute $V_k=Y_k^{(1)}-Y_k^{(2)}$. Compute the sample means and sample variances of U's and V's, respectively, obtaining $\bar{U}$ and S_U^2 for U's, $\bar{V}$ and S_V^2 for V's.	(1) Suppose N=41 and M=70. After computing the U's and V's, and their sample means and variances, one gets $\bar{U} = 0.695$, $S_U^2 = 0.146$ $\bar{V} = 0.399$, and $S_V^2 = 0.106$
(2) Let σ_U^2 and σ_V^2 be the variances of U and V, respectively. Test H_0: $\sigma_U^2 = \sigma_V^2$ by the variance ratio test: If $S_U^2 > S_V^2$, form $F_0 = \frac{S_U^2}{S_V^2}$, where S_U^2 and S_V^2 are obtained in (1). Under H_0, F_0	(2) Test H_0: $\sigma_U^2 = \sigma_V^2$. Form $F_0 = \frac{(0.383)^2}{(0.326)^2} = 1.380$. From the table for percentile of the F distribution (Table A-5, Appendix A), find $p=F_{.95}(69,40) \equiv 1.6$. Since $p > F_0$, H_0 is accepted.

has an F distribution with N-1 and M-1 degrees of freedom. Find the value $p=F_{.95}(N-1,M-1)$ from the table for percentiles of the F distribution (Table A-5, Appendix A). If F_0 is smaller than p, accept H_0; otherwise, reject H_0.

(3) If H_0 in (2) is accepted, then use the two-sample t test for testing $H_0^{(1)}$. Otherwise, go to (4). The test statistic is

$$t_o = \frac{\bar{U} - \bar{V}}{S_p(\frac{1}{N} + \frac{1}{M})^{1/2}}$$

where $S_p^2 = [(N-1)S_U^2 + (M-1)S_V^2]/N+M-2$.

Under $H_0^{(1)}$, t_0 has a student's t distribution with N+M-2 degrees of freedom. Find p, the 95th percentile of the student's t distribution with N+M-2 degrees of freedom (Table A-4). If t_0 is smaller than p, accept $H_0^{(1)}$ at the $\alpha=0.05$ level. Otherwise, reject $H_0^{(1)}$.

(3) Since H_0 is accepted in (2), the two-sample t test can be used for testing $H_0^{(1)}$. Compute

$$S_p^2 = \frac{(41-1)(0.146)+(70-1)(0.106)}{41 + 70 - 2}$$

$$= 0.132$$

The statistic is

$$t_o = \frac{0.695 - 0.399}{\sqrt{0.132(\frac{1}{41} + \frac{1}{70})}} = 4.14$$

with 109 degrees of freedom. Since the p value is 1.658, which is less than t_0, $H_0^{(1)}$ is rejected at $\alpha=0.05$ level. In other words, at time 1 and time 2, the difference between Portland and Eugene

(4) If H_0 in (2) is rejected, then use the Welch two-sample test for testing $H_0^{(1)}$. The test statistic is

$$t_0 = \frac{\overline{U} - \overline{V}}{\sqrt{\frac{S_U^2}{N} + \frac{S_V^2}{M}}} .$$

Under $H_0^{(1)}$, t_0 has an approximate student's t distribution with degrees of freedom of

$$\nu = \frac{\left(\frac{S_U^2}{N} + \frac{S_V^2}{M}\right)^2}{\frac{S_U^4}{N^2(N-1)} + \frac{S_V^4}{M^2(M-1)}}$$

Find p, the value of the 95th percentile of the student's t distribution with ν degrees of freedom. If t_0 is smaller than p, accept $H_0^{(1)}$ at α=0.05 level; otherwise, reject $H_0^{(1)}$.

emission slopes is significant at α=0.05 level.

(4) In the case where σ_u^2 and σ_v^2 are shown to be quite different according to (2), the Welch two-sample test is used. Thus, compute

$$t_0 = \frac{0.695 - 0.399}{\sqrt{\frac{0.146}{41} + \frac{0.106}{70}}} = 0.125$$

with degrees of freedom

$$\nu = \frac{\left(\frac{0.146}{41} + \frac{0.106}{70}\right)^2}{\frac{(0.146)^2}{41^2 \cdot 40} + \frac{(0.106)^2}{70^2 \cdot 69}}$$

$$\cong 40$$

Since the p value is 1.684, which is greater than t_0, $H_0^{(1)}$ is accepted at α=0.05 level. In other words, at time 1 and time 2, the difference between Portland and Eugene emission slopes is not significant at α=0.05 level.

The same procedure can be used for testing the difference between slopes for passed versus failed vehicles.

Method 2. Perform an analysis of covariance according to the technique described in Section 3.3.3 with any one pollutant-specific variable as the dependent variable, region (Portland or Eugene) as independent variable, and time as the covariate. Test for equality of slopes among the two regions according to (2) in the Interpretation of Output, Section 3.3.3. The acceptance of the equality of slopes means that there is no significant difference between the Portland and Eugene slopes.

3.3.1.3 Can idle/hot-start FTP and FTP deterioration be related by equations? The answer relies on regression analysis. The analysis is broken into several steps:

Step 1. Choose a pollutant-specific FTP variable, say, Y, and one of the idle or hot-start FTP variables, say, X.

Step 2. Suppose there are N vehicles. Let Y_{ij} be the pollutant value for the i^{th} car measured at j^{th} time, t_j, where $i=1,\ldots,N$ and $j=1,2,3,4$. For the i^{th} vehicle, compute the regression coefficient, β_i, as

$$\beta_i = \left[4\sum_{j=1}^{4} t_j Y_{ij} - \left(\sum_{j=1}^{4} t_j\right)\left(\sum_{j=1}^{4} Y_{ij}\right)\right] \Big/ \left[4\left(\sum_{j=1}^{4} t_j^2\right) - \left(\sum_{j=1}^{4} t_j\right)^2\right] .$$

β_i can be considered as the "pollutant-specific FTP deterioration" for the i^{th} vehicle, where $i=1,\ldots,N$.

Step 3. Let β be the pollutant-specific FTP deterioration. Perform a linear regression analysis with β as dependent variable and X as independent variable according to the procedure given in Section 3.2.5. Let b_1

be the regression coefficient found in the output.

Step 4. Perform a linear regression analysis with X as the dependent variable and β as the independent variable according to the same procedure used in Step 2. Let b_2 be the regression coefficient found in the output.

Step 5. Compute $\rho^2 = b_1 \cdot b_2$, the correlation coefficient, where b_1 and b_2 are from Step 3 and Step 4, respectively. ρ^2 always lies between 0 and 1. ρ^2 achieves 1 if and only if β and X are in strict linear functional relationship. If and only if $\rho=0$, where X and ρ are said to be uncorrelated, are the two regression lines obtained in Steps 3 and 4 at right angles to each other.

3.3.2 Does fuel economy seem to deteriorate with CO_2 deterioration?

Regression analysis can be used following the steps shown below.

Step 1. Suppose there are N vehicles. Let Y_{ij} be the fuel economy of the i^{th} vehicle measured at j^{th} time, t_j, where $i=1,\ldots,N$, and $j=1,2,3,4$. Compute the fuel economy deterioration for the i^{th} vehicle, β_i, according to the formula given in Section 3.3.1.3, Step 2, where $i=1,\ldots,N$. Similarly, let Z_{ij} be the CO_2 value of the i^{th} vehicle measured at time t_j, where $i=1,\ldots,N$, $j=1,2,3,4$, and compute the CO_2 deterioration for the i^{th} car, α_i, according to the formula in Section 3.3.1.3, Step 2, where $i=1,\ldots,N$.

Step 2. Let β be the fuel economy deterioration and α be the CO_2 deterioration. Perform a linear regression analysis with β as dependent variable and α as independent variable, according to the procedure given in Section 3.2.5. Let b_1 be the regression coefficient found in the output.

Step 3. Perform a linear regression analysis with α as dependent variable and β as independent variable according to the procedure given in Section 3.2.5. Let b_2 be the regression coefficient found in the output.

Step 4. Compute $\rho^2 = b_1 \cdot b_2$, the correlation coefficient, where b_1 and b_2 are from Step 2 and Step 3, respectively. ρ^2 always lies between 0 and 1. ρ^2 achieves the value 1 if and only if α and β are in a strict linear functional relationship. If and only if $\rho = 0$, where α and β are said to be uncorrelated, are the two regression lines obtained in Steps 3 and 4 at right angles to each other.

3.3.3 Is there a seasonal trend, in temperature, to deterioration on the FTP? on the hot-start FTP? on the idle test? on the HFET?

The one-way analysis-of-covariance technique can be used in this problem. The details of the technique, together with the computer subroutine using FTP as an example, are given below.

ONE-WAY ANALYSIS OF COVARIANCE

(1) Select primary variables, including dependent, independent and covariate.

EXAMPLE: The dependent variable is the pollutant-specific FTP emission variable (say, CO). The independent variable is TIME (3 levels). The covariate is TEMP (continuous variable).

(2) Form the table $\{y_{ij}, x_{ij}\}$, where y_{ij}, x_{ij} are the observed values of the j^{th}-experiment unit in the i^{th} cell.

EXAMPLE: Form the table $\{y_{ij}, x_{ij}\}$, where y_{ij} is the observed FTP CO emission value for the j^{th} car measured at time i, and x_{ij} is the corresponding temperature.

(3) Consider the analysis-of-covariance model

$$Y_{ij} = \mu + \alpha_i + \beta(x_{ij}) + \ell_{ij} \quad .$$

Interpretation of the model:

μ	is the overall mean
α_i	is the main effect of time i
β	is the regression coefficient for temperature
ℓ_{ij}	is the error term .

Here, we are interested in testing the hypothesis $H_0{:}\beta=0$. To reject the hypothesis means there is a seasonal trend, in temperature, to deterioration on FTP.

(4) Most of the statistical package programs contain a one-way analysis-of-covariance subroutine, e.g., SPSS-ANOVA AND ONE WAY, and BMDP1V. In the following, the output of BMDP1V is used as an illustration.

ONE-WAY ANALYSIS OF COVARIANCE (BMDP1V)

For program card preparation, one should consult the BMDP manual. Special instructions:

(i) GROUPING IS TIME
(ii) DEPENDENT IS CO
(iii) INDEPENDENT IS TEMP.

INTERPRETATION OF THE OUTPUT EXAMPLE

PROCEDURE	EXAMPLE
(1) Look for the Analysis of Variance table in the output of BMDP1V.	(1) Look for the Analysis of Variance table in the OUTPUT example.
(2) Look for the TAIL AREA PROB. of the Equality of Slopes. If the	(2) The TAIL AREA PROB. of the Equality of Slopes is 0.907, which

OUTPUT EXAMPLE (BMDP1V)

•
•
•

```
Estimates of Means

                             TIME                      TOTAL
                        1          2           3            4
TEMP          2       2.000      3.750       1.750        2.500
   CO         3       9.250     11.750      12.500        9.166
   COVARIATE      REG. COEFF.        STD. ERR.        T-VALUE
   TEMP             0.833              0.433           1.920


   GROUP          GRP. MEAN        ADJ.GRP.MEAN      STD.ERR.
     1              9.250              9.666           0.632
     2             11.750             10.708           0.804
     3              6.500              7.125           0.677
Analysis of Variance

Source of Variance           D.F.   SUM OF SQ.  MEAN SQ.  F-VALUE   TAIL AREA
                                                                      PROB.

Equality of Adj.Cell Means    2      17.492      8.746     6.196     0.237
Zero Slope                    1       5.208      5.208     3.690     0.091
  Error                       8      11.291      1.411

Equality of Slopes            2       0.359      0.179     0.098     0.907
  Error                       6      10.931      1.822


Slope within Each Group

                              TIME
                        1      2      3

     TEMP      2       1.00   1.00   0.54
```

INTERPRETATION OF THE OUTPUT EXAMPLE (Cont'd)

value is greater than 0.05, then the analysis of covariance model (3) (above) holds. Otherwise, the three slopes are different and the model cannot be assumed.

(3) If the analysis of covariance model is confirmed in (2), look for the TAIL AREA PROB. in the Zero Slope (source of variance). If the value is greater than 0.05, then the covariance regression coefficient is significant, and there is a significant seasonal trend, in temperature, to the FTP CO-values. If the value is less than 0.05, then the regression coefficient is nonsignificant and there exists no seasonal trend. If the analysis of covariance model is not valid as shown in (2), look at the Slope within Each Group table, which gives an idea of the relative magnitude of the slopes in each time cell. No further conclusion can be drawn here.

is greater than 0.05. Thus the analysis-of-covariance model of (3) (above) holds.

(3) From (2), the analysis of covariance model is confirmed. Thus look for the TAIL AREA PROB. in the Zero Slope (source of variance) and find the value is 0.091. This value is greater than 0.05; therefore, there is a significant seasonal trend to the FTP-CO value.

INTERPRETATION OF THE OUTPUT EXAMPLE (Cont'd)

(4) If, from (3), it has been confirmed that the slope is not zero, look for the table of Estimates of Means and find the regression coefficient and standard error of the covariate, which will give one an idea about the magnitude of the slope.

(5) The CO deterioration can be seen from the Estimates of Mean table.

(4) From (3), it has been confirmed that the slope is not zero. Look for the table of Estimates of Mean and find REG. COEFF = 0.833, STD.ERR= 0.433, which gives one an impression of the magnitude of the seasonal trend.

(5) The CO deterioration can be seen from the Estimates of Mean table:

	TIME		
	1	2	3
CO	9.250	11.750	12.500

One concludes that CO deterioration does exist and has a seasonal trend.

3.3.4 Is there a relationship between maintenance performed (a categorical variable) and rate of ST, FTP, and fuel economy deterioration?

Apply the AID algorithm described in Section 3.2.7 with emission value as the dependent variable and CATEGORY and TIME as the independent variables.

3.3.5 How long does it take a failed and subsequently maintained vehicle to reach its pre-maintenance ST/FTP/fuel economy levels?

The answer to this question relies on a properly designed study. One method has the following procedure:

PROCEDURE

(1) Take a random sample of size n vehicles from among the set of failed and subsequently maintained vehicles.

(2) Once every fixed-time interval, say, once every three months, check the vehicle's ST/FTP/fuel economy levels. Let $T_0 < T_1 < \ldots < T_M$ (here, M can be set equal to 3) be the scheduled checking times. Then follow each car throughout the time interval from T_0 to T_M.

(3) Let X be an emission/fuel economy variable. Let X_{ij} be the observed X value for the i^{th} car at time T_j, where i=1,...,N, and j=1,...,M. For each car i, compute

$$\beta_i = \left[M\sum_{j=1}^{M} t_j X_{ij} - \left(\sum_{j=1}^{M} t_j\right)\left(\sum_{j=1}^{M} X_{ij}\right)\right] \Bigg/ \left[M\left(\sum_{j=1}^{M} t_j^2\right) - \left(\sum_{j=1}^{M} t_j\right)^2\right] \quad ,$$

$$\overline{X} = \left(\sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij}\right) \Bigg/ (NM) \quad ,$$

$$\overline{T} = \left(\sum_{j=1}^{M} T_j\right) \Bigg/ M \quad ,$$

and

$$\alpha_i = \overline{X} - \beta_i \overline{T} \quad ,$$

where i=1,...,N.

NOTE: α_i and β_i are the coefficients of the linear regression of X on T.

(4) Let ν_i be the pre-maintenance level for the i^{th} car, where i=1,...,N. Compute

$$t_i = \frac{\nu_i - \alpha_i}{\beta_i} \quad , \qquad \text{where } i=1,\ldots,N \quad .$$

NOTE: t_i is the period for the i^{th} tailed and subsequently maintained vehicle to reach its pre-maintenance level.

(5) Compute the sample mean, $\bar{t}$, and sample standard deviation, $S_d(t)$, of t_i, where i=1,...,N. Then $\bar{t}$ gives the average time to take a failed and subsequently maintained vehicle to reach its pre-maintenance ST/FTP/fuel economy levels. The 95% confidence interval is given by the two limits

$$\bar{t} \pm 1.96 \, S_d(t) \quad .$$

3.3.6 Is the average amount of money spent on maintenance between the initial retest and the final test different from Portland to Eugene by model-year group? By pass/fail group?

The two-sample t test or Welsh test mentioned in 3.3.1.2, Method 1, can be used. Specifically, suppose one wants to test whether the average amount of money spent on maintenance differs from Portland to Eugene. Let U_j be the amount of money spent on maintenance for the j^{th} car in the 1972-1974 Portland vehicles, j=1,...,N. Let V_k be that money spent on maintenance for the k^{th} car in the 1972-1974 Eugene cars. Compute the sample means and sample variances of U's and V's, respectively, and then follow the procedure in 3.3.1.2, Method 1.

APPENDIX A

STATISTICAL TABLES

Table A-1a Table for Testing Skewness

(One-tailed percentage points of the distribution of $\sqrt{b_1} = g_1 = m_3/m_2^{3/2}$)*

Size of Sample n	Percentage Points 5%	1%	Standard Deviation	Size of Sample n	Percentage Points 5%	1%	Standard Deviation
25	0.711	1.061	0.4354	100	0.389	0.567	0.2377
30	0.662	0.986	.4052	125	0.350	0.508	.2139
35	0.621	0.923	.3804	150	0.321	0.464	.1961
40	0.587	0.870	.3596	175	0.298	0.430	.1820
45	0.558	0.825	.3418	200	0.280	0.403	.1706
50	0.534	0.787	.3264				
				250	0.251	0.360	.1531
60	0.492	0.723	.3009	300	0.230	0.329	.1400
70	0.459	0.673	.2806	350	0.213	0.305	.1298
80	0.432	0.631	.2638	400	0.200	0.285	.1216
90	0.409	0.596	.2498	450	0.188	0.269	.1147
100	0.389	0.567	.2377	500	0.179	0.255	.1089

* Since the distribution of $\sqrt{b_1}$ is symmetrical about zero, the percentage points represent 10% and 2% two-tailed values.

Table A-1b Table for Testing Kurtosis

(Percentage points of the distribution of $b_2 = m_4/m_2^2$)*

Size of Sample n	Percentage Points Upper 1%	Upper 5%	Lower 5%	Lower 1%	Size of Sample n	Percentage Points Upper 1%	Upper 5%	Lower 5%	Lower 1%
50	4.88	3.99	2.15	1.95	600	3.54	3.34	2.70	2.60
75	4.59	3.87	2.27	2.08	650	3.52	3.33	2.71	2.61
100	4.39	3.77	2.35	2.18	700	3.50	3.31	2.72	2.62
125	4.24	3.71	2.40	2.24	750	3.48	3.30	2.73	2.64
150	4.13	3.65	2.45	2.29	800	3.46	3.29	2.74	2.65
					850	3.45	3.28	2.74	2.66
200	3.98	3.57	2.51	2.37	900	3.43	3.28	2.75	2.66
250	3.87	3.52	2.55	2.42	950	3.42	3.27	2.76	2.67
300	3.79	3.47	2.59	2.46	1000	3.41	3.26	2.76	2.68
350	3.72	3.44	2.62	2.50					
400	3.67	3.41	2.64	2.52	1200	3.37	3.24	2.78	2.71
450	3.63	3.39	2.66	2.55	1400	3.34	3.22	2.80	2.72
500	3.60	3.37	2.67	2.57	1600	3.32	3.21	2.81	2.74
550	3.57	3.35	2.69	2.58	1800	3.30	3.20	2.82	2.76
600	3.54	3.34	2.70	2.60	2000	3.28	3.18	2.83	2.77

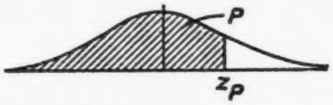
Table A-2a Critical Values for the Kolmogorov-Smirnov Test of Goodness-of-Fit

Sample Size (n)	Significance Level .20	.15	.10	.05	.01
1	.900	.925	.950	.975	.995
2	.684	.726	.776	.842	.929
3	.565	.597	.642	.708	.829
4	.494	.525	.564	.624	.734
5	.446	.474	.510	.563	.669
6	.410	.436	.470	.521	.618
7	.381	.405	.438	.486	.577
8	.358	.381	.411	.457	.543
9	.339	.360	.388	.432	.514
10	.322	.342	.368	.409	.486
11	.307	.326	.352	.391	.468
12	.295	.313	.338	.375	.450
13	.284	.302	.325	.361	.433
14	.274	.292	.314	.349	.418
15	.266	.283	.304	.338	.404
16	.258	.274	.295	.328	.391
17	.250	.266	.286	.318	.380
18	.244	.259	.278	.309	.370
19	.237	.252	.272	.301	.361
20	.231	.246	.264	.294	.352
25	.21	.22	.24	.264	.32
30	.19	.20	.22	.242	.29
35	.18	.19	.21	.23	.27
40				.21	.25
50				.19	.23
60				.17	.21
70				.16	.19
80				.15	.18
90				.14	
100				.14	
Asymptotic Formula:	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Table A-2b Critical Values for the Kolmogorov-Smirnov Test of Goodness-of-Fit (use sample mean and standard deviation)

Sample Size N	Level of Significance for $D = \mathrm{Max}\ \lvert F^*(X) - S_N(X) \rvert$				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.149	.153	.165	.180	.203
30	.131	.136	.144	.161	.187
Over 30	$\frac{.736}{\sqrt{N}}$	$\frac{.768}{\sqrt{N}}$	$\frac{.805}{\sqrt{N}}$	$\frac{.886}{\sqrt{N}}$	$\frac{1.031}{\sqrt{N}}$

Table A-3 Cumulative Normal Distribution--Values of P



(Values of P corresponding to z_p for the normal curve. z is the standard normal variable. The value for P for $-z_p$ equals one minus the value of P for $+z_p$, e.g., the P for -1.62 equals 1 - .9474 = .0526.)

z_p	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table A-4 Percentiles of the Student's t Distribution

$\nu \backslash q$	60	75	90	95	97.5	99	99.5	99.95
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.289	.816	1.886	2.920	4.303	6.965	9.925	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	12.941
4	.271	.741	1.533	2.132	2.776	3.747	4.604	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	6.859
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	5.405
8	.262	.706	2.397	1.860	2.306	2.896	3.355	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	4.140
15	.258	691	1.341	1.753	2.131	2.602	2.947	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.291

Table A-5 Percentiles of the F-Distribution

95th Percentile

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	236·8	238·9	240·5	241·9	243·9	245·9	248·0	249·1	250·1	251·1	252·2	253·3	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·35	19·37	19·38	19·40	19·41	19·43	19·45	19·45	19·46	19·47	19·48	19·49	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·89	8·85	8·81	8·79	8·74	8·70	8·66	8·64	8·62	8·59	8·57	8·55	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·09	6·04	6·00	5·96	5·91	5·86	5·80	5·77	5·75	5·72	5·69	5·66	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·88	4·82	4·77	4·74	4·68	4·62	4·56	4·53	4·50	4·46	4·43	4·40	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·21	4·15	4·10	4·06	4·00	3·94	3·87	3·84	3·81	3·77	3·74	3·70	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·79	3·73	3·68	3·64	3·57	3·51	3·44	3·41	3·38	3·34	3·30	3·27	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·50	3·44	3·39	3·35	3·28	3·22	3·15	3·12	3·08	3·04	3·01	2·97	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·29	3·23	3·18	3·14	3·07	3·01	2·94	2·90	2·86	2·83	2·79	2·75	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·14	3·07	3·02	2·98	2·91	2·85	2·77	2·74	2·70	2·66	2·62	2·58	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	3·01	2·95	2·90	2·85	2·79	2·72	2·65	2·61	2·57	2·53	2·49	2·45	2·40
12	4·75	3·89	3·49	3·26	3·11	3·00	2·91	2·85	2·80	2·75	2·69	2·62	2·54	2·51	2·47	2·43	2·38	2·34	2·30
13	4·67	3·81	3·41	3·18	3·03	2·92	2·83	2·77	2·71	2·67	2·60	2·53	2·46	2·42	2·38	2·34	2·30	2·25	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·76	2·70	2·65	2·60	2·53	2·46	2·39	2·35	2·31	2·27	2·22	2·18	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·71	2·64	2·59	2·54	2·48	2·40	2·33	2·29	2·25	2·20	2·16	2·11	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·66	2·59	2·54	2·49	2·42	2·35	2·28	2·24	2·19	2·15	2·11	2·06	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·61	2·55	2·49	2·45	2·38	2·31	2·23	2·19	2·15	2·10	2·06	2·01	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·58	2·51	2·46	2·41	2·34	2·27	2·19	2·15	2·11	2·06	2·02	1·97	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·54	2·48	2·42	2·38	2·31	2·23	2·16	2·11	2·07	2·03	1·98	1·93	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·51	2·45	2·39	2·35	2·28	2·20	2·12	2·08	2·04	1·99	1·95	1·90	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·49	2·42	2·37	2·32	2·25	2·18	2·10	2·05	2·01	1·96	1·92	1·87	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·46	2·40	2·34	2·30	2·23	2·15	2·07	2·03	1·98	1·94	1·89	1·84	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·44	2·37	2·32	2·27	2·20	2·13	2·05	2·01	1·96	1·91	1·86	1·81	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·42	2·36	2·30	2·25	2·18	2·11	2·03	1·98	1·94	1·89	1·84	1·79	1·73
25	4·24	3·39	2·99	2·76	2·60	2·49	2·40	2·34	2·28	2·24	2·16	2·09	2·01	1·96	1·92	1·87	1·82	1·77	1·71
26	4·23	3·37	2·98	2·74	2·59	2·47	2·39	2·32	2·27	2·22	2·15	2·07	1·99	1·95	1·90	1·85	1·80	1·75	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·37	2·31	2·25	2·20	2·13	2·06	1·97	1·93	1·88	1·84	1·79	1·73	1·67
28	4·20	3·34	2·95	2·71	2·56	2·45	2·36	2·29	2·24	2·19	2·12	2·04	1·96	1·91	1·87	1·82	1·77	1·71	1·65
29	4·18	3·33	2·93	2·70	2·55	2·43	2·35	2·28	2·22	2·18	2·10	2·03	1·94	1·90	1·85	1·81	1·75	1·70	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·33	2·27	2·21	2·16	2·09	2·01	1·93	1·89	1·84	1·79	1·74	1·68	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·25	2·18	2·12	2·08	2·00	1·92	1·84	1·79	1·74	1·69	1·64	1·58	1·51
60	4·00	3·15	2·76	2·53	2·37	2·25	2·17	2·10	2·04	1·99	1·92	1·84	1·75	1·70	1·65	1·59	1·53	1·47	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·09	2·02	1·96	1·91	1·83	1·75	1·66	1·61	1·55	1·50	1·43	1·35	1·25
∞	3·84	3·00	2·60	2·37	2·21	2·10	2·01	1·94	1·88	1·83	1·75	1·67	1·57	1·52	1·46	1·39	1·32	1·22	1·00

Table A-5 - Cont'd

99th Percentile

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999·5	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98·50	99·00	99·17	99·25	99·30	99·33	99·36	99·37	99·39	99·40	99·42	99·43	99·45	99·46	99·47	99·47	99·48	99·49	99·50
3	34·12	30·82	29·46	28·71	28·24	27·91	27·67	27·49	27·35	27·23	27·05	26·87	26·69	26·60	26·50	26·41	26·32	26·22	26·13
4	21·20	18·00	16·69	15·98	15·52	15·21	14·98	14·80	14·66	14·55	14·37	14·20	14·02	13·93	13·84	13·75	13·65	13·56	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·46	10·29	10·16	10·05	9·89	9·72	9·55	9·47	9·38	9·29	9·20	9·11	9·02
6	13·75	10·92	9·78	9·15	8·75	8·47	8·26	8·10	7·98	7·87	7·72	7·56	7·40	7·31	7·23	7·14	7·06	6·97	6·88
7	12·25	9·55	8·45	7·85	7·46	7·19	6·99	6·84	6·72	6·62	6·47	6·31	6·16	6·07	5·99	5·91	5·82	5·74	5·65
8	11·26	8·65	7·59	7·01	6·63	6·37	6·18	6·03	5·91	5·81	5·67	5·52	5·36	5·28	5·20	5·12	5·03	4·95	4·86
9	10·56	8·02	6·99	6·42	6·06	5·80	5·61	5·47	5·35	5·26	5·11	4·96	4·81	4·73	4·65	4·57	4·48	4·40	4·31
10	10·04	7·56	6·55	5·99	5·64	5·39	5·20	5·06	4·94	4·85	4·71	4·56	4·41	4·33	4·25	4·17	4·08	4·00	3·91
11	9·65	7·21	6·22	5·67	5·32	5·07	4·89	4·74	4·63	4·54	4·40	4·25	4·10	4·02	3·94	3·86	3·78	3·69	3·60
12	9·33	6·93	5·95	5·41	5·06	4·82	4·64	4·50	4·39	4·30	4·16	4·01	3·86	3·78	3·70	3·62	3·54	3·45	3·36
13	9·07	6·70	5·74	5·21	4·86	4·62	4·44	4·30	4·19	4·10	3·96	3·82	3·66	3·59	3·51	3·43	3·34	3·25	3·17
14	8·86	6·51	5·56	5·04	4·69	4·46	4·28	4·14	4·03	3·94	3·80	3·66	3·51	3·43	3·35	3·27	3·18	3·09	3·00
15	8·68	6·36	5·42	4·89	4·56	4·32	4·14	4·00	3·89	3·80	3·67	3·52	3·37	3·29	3·21	3·13	3·05	2·96	2·87
16	8·53	6·23	5·29	4·77	4·44	4·20	4·03	3·89	3·78	3·69	3·55	3·41	3·26	3·18	3·10	3·02	2·93	2·84	2·75
17	8·40	6·11	5·18	4·67	4·34	4·10	3·93	3·79	3·68	3·59	3·46	3·31	3·16	3·08	3·00	2·92	2·83	2·75	2·65
18	8·29	6·01	5·09	4·58	4·25	4·01	3·84	3·71	3·60	3·51	3·37	3·23	3·08	3·00	2·92	2·84	2·75	2·66	2·57
19	8·18	5·93	5·01	4·50	4·17	3·94	3·77	3·63	3·52	3·43	3·30	3·15	3·00	2·92	2·84	2·76	2·67	2·58	2·49
20	8·10	5·85	4·94	4·43	4·10	3·87	3·70	3·56	3·46	3·37	3·23	3·09	2·94	2·86	2·78	2·69	2·61	2·52	2·42
21	8·02	5·78	4·87	4·37	4·04	3·81	3·64	3·51	3·40	3·31	3·17	3·03	2·88	2·80	2·72	2·64	2·55	2·46	2·36
22	7·95	5·72	4·82	4·31	3·99	3·76	3·59	3·45	3·35	3·26	3·12	2·98	2·83	2·75	2·67	2·58	2·50	2·40	2·31
23	7·88	5·66	4·76	4·26	3·94	3·71	3·54	3·41	3·30	3·21	3·07	2·93	2·78	2·70	2·62	2·54	2·45	2·35	2·26
24	7·82	5·61	4·72	4·22	3·90	3·67	3·50	3·36	3·26	3·17	3·03	2·89	2·74	2·66	2·58	2·49	2·40	2·31	2·21
25	7·77	5·57	4·68	4·18	3·85	3·63	3·46	3·32	3·22	3·13	2·99	2·85	2·70	2·62	2·54	2·45	2·36	2·27	2·17
26	7·72	5·53	4·64	4·14	3·82	3·59	3·42	3·29	3·18	3·09	2·96	2·81	2·66	2·58	2·50	2·42	2·33	2·23	2·13
27	7·68	5·49	4·60	4·11	3·78	3·56	3·39	3·26	3·15	3·06	2·93	2·78	2·63	2·55	2·47	2·38	2·29	2·20	2·10
28	7·64	5·45	4·57	4·07	3·75	3·53	3·36	3·23	3·12	3·03	2·90	2·75	2·60	2·52	2·44	2·35	2·26	2·17	2·06
29	7·60	5·42	4·54	4·04	3·73	3·50	3·33	3·20	3·09	3·00	2·87	2·73	2·57	2·49	2·41	2·33	2·23	2·14	2·03
30	7·56	5·39	4·51	4·02	3·70	3·47	3·30	3·17	3·07	2·98	2·84	2·70	2·55	2·47	2·39	2·30	2·21	2·11	2·01
40	7·31	5·18	4·31	3·83	3·51	3·29	3·12	2·99	2·89	2·80	2·66	2·52	2·37	2·29	2·20	2·11	2·02	1·92	1·80
60	7·08	4·98	4·13	3·65	3·34	3·12	2·95	2·82	2·72	2·63	2·50	2·35	2·20	2·12	2·03	1·94	1·84	1·73	1·60
120	6·85	4·79	3·95	3·48	3·17	2·96	2·79	2·66	2·56	2·47	2·34	2·19	2·03	1·95	1·86	1·76	1·66	1·53	1·38
∞	6·63	4·61	3·78	3·32	3·02	2·80	2·64	2·51	2·41	2·32	2·18	2·04	1·88	1·79	1·70	1·59	1·47	1·32	1·00

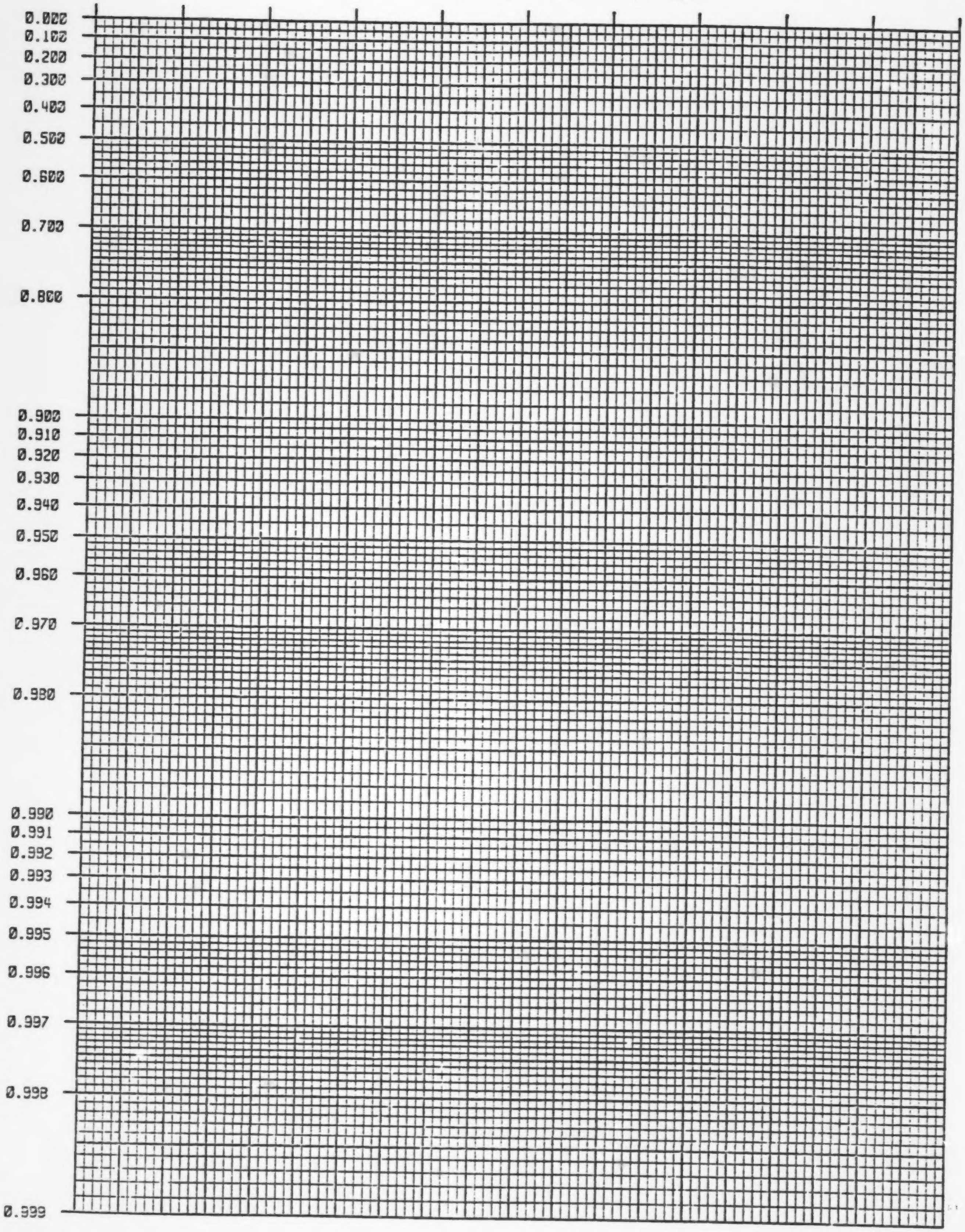
Table A-6 Percentiles of the χ^2 Distribution

ν \ q	0.5	1	2.5	5	10	20	30	40
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	0.0158	0.0642	0.148	0.275
2	0.0100	0.0201	0.0506	0.103	0.211	0.446	0.713	1.02
3	0.0717	0.115	0.216	0.352	0.584	1.00	1.42	1.87
4	0.207	0.297	0.484	0.711	1.06	1.65	2.19	2.75
5	0.412	0.554	0.831	1.15	1.61	2.34	3.00	3.66
6	0.676	0.872	1.24	1.64	2.20	3.07	3.83	4.57
7	0.989	1.24	1.69	2.17	2.83	3.82	4.67	5.49
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	6.42
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	7.36
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	8.30
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	9.24
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	10.2
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	11.1
14	4.07	4.66	5.63	6.57	7.79	9.47	10.8	12.1
15	4.60	5.23	6.26	7.26	8.55	10.3	11.7	13.0
16	5.14	5.81	6.91	7.96	9.31	11.2	12.6	14.0
17	5.70	6.41	7.56	8.67	10.1	12.0	13.5	14.9
18	6.26	7.01	8.23	9.39	10.9	12.9	14.4	15.9
19	6.84	7.63	8.91	10.1	11.7	13.7	15.4	16.9
20	7.43	8.26	9.59	10.9	12.4	14.6	16.3	17.8
21	8.03	8.90	10.3	11.6	13.2	15.4	17.2	18.8
22	8.64	9.54	11.0	12.3	14.0	16.3	18.1	19.7
23	9.26	10.2	11.7	13.1	14.8	17.2	19.0	20.7
24	9.89	10.9	12.4	13.8	15.7	18.1	19.9	21.7
25	10.5	11.5	13.1	14.6	16.5	18.9	20.9	22.6
26	11.2	12.2	13.8	15.4	17.3	19.8	21.8	23.6
27	11.8	12.9	14.6	16.2	18.1	20.7	22.7	24.5
28	12.5	13.6	15.3	16.9	18.9	21.6	23.6	25.5
29	13.1	14.3	16.0	17.7	19.8	22.5	24.6	26.5
30	13.8	15.0	16.8	18.5	20.6	23.4	25.5	27.4
35	17.2	18.5	20.6	22.5	24.8	27.8	30.2	32.3
40	20.7	22.2	24.4	26.5	29.1	32.3	34.9	37.1
45	24.3	25.9	28.4	30.6	33.4	36.9	39.6	42.0
50	28.0	29.7	32.4	34.8	37.7	41.4	44.3	46.9
75	47.2	49.5	52.9	56.1	59.8	64.5	68.1	71.3
100	67.3	70.1	74.2	77.9	82.4	87.9	92.1	95.8

Table A-6 - Cont'd

ν \ q	50	60	70	80	90	95	97.5	99	99.5	99.9
1	0.455	0.708	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.8
2	1.39	1.83	2.41	3.22	4.61	5.99	7.38	9.21	10.6	13.8
3	2.37	2.95	3.67	4.64	6.25	7.81	9.35	11.3	12.8	16.3
4	3.36	4.04	4.88	5.99	7.78	9.49	11.1	13.3	14.9	18.5
5	4.35	5.13	6.06	7.29	9.24	11.1	12.8	15.1	16.7	20.5
6	5.35	6.21	7.23	8.56	10.6	12.6	14.4	16.8	18.5	22.5
7	6.35	7.28	8.38	9.80	12.0	14.1	16.0	18.5	20.3	24.3
8	7.34	8.35	9.52	11.0	13.4	15.5	17.5	20.1	22.0	26.1
9	8.34	9.41	10.7	12.2	14.7	16.9	19.0	21.7	23.6	27.9
10	9.34	10.5	11.8	13.4	16.0	18.3	20.5	23.2	25.2	29.6
11	10.3	11.5	12.9	14.6	17.3	19.7	21.9	24.7	26.8	31.3
12	11.3	12.6	14.0	15.8	18.5	21.0	23.3	26.2	28.3	32.9
13	12.3	13.6	15.1	17.0	19.8	22.4	24.7	27.7	29.8	34.5
14	13.3	14.7	16.2	18.2	21.1	23.7	26.1	29.1	31.3	36.1
15	14.3	15.7	17.3	19.3	22.3	25.0	27.5	30.6	32.8	37.7
16	15.3	16.8	18.4	20.5	23.5	26.3	28.8	32.0	34.3	39.3
17	16.3	17.8	19.5	21.6	24.8	27.6	30.2	33.4	35.7	40.8
18	17.3	18.9	20.6	22.8	26.0	28.9	31.5	34.8	37.2	42.3
19	18.3	19.9	21.7	23.9	27.2	30.1	32.9	36.2	38.6	43.8
20	19.3	21.0	22.8	25.0	28.4	31.4	34.2	37.6	40.0	45.3
21	20.3	22.0	23.9	26.9	29.6	32.7	35.5	38.9	41.4	46.8
22	21.3	23.0	24.9	27.3	30.8	33.9	36.8	40.3	42.8	48.3
23	22.3	24.1	26.0	28.4	32.0	35.2	38.1	41.6	44.2	49.7
24	23.3	25.1	27.1	29.6	33.2	36.4	39.4	43.0	45.6	51.2
25	24.3	26.1	28.2	30.7	34.4	37.7	40.6	44.3	46.9	52.6
26	25.3	27.2	29.2	31.8	35.6	38.9	41.9	45.6	48.3	54.1
27	26.3	28.2	30.3	32.9	36.7	40.1	43.2	47.0	49.6	55.5
28	27.3	29.2	31.4	34.0	37.9	41.3	44.5	48.3	51.0	56.9
29	28.3	30.3	32.5	35.1	39.1	42.6	45.7	49.6	52.3	58.3
30	29.3	31.3	33.5	36.3	40.3	43.8	47.0	50.9	53.7	59.7
35	34.3	36.5	38.9	41.8	46.1	49.8	53.2	57.3	60.3	66.6
40	39.3	41.6	44.2	47.3	51.8	55.8	59.3	63.7	66.8	73.4
45	44.3	46.8	49.5	52.7	57.5	61.7	65.4	70.0	73.2	80.1
50	49.3	51.9	54.7	58.2	63.2	67.5	71.4	76.2	79.5	86.7
75	74.3	77.5	80.9	85.1	91.1	96.2	100.8	106.4	110.3	118.6
100	99.3	102.9	106.9	111.7	118.5	124.3	129.6	135.6	140.2	149.4

Table A-7 Sample of Semilog Probability Paper



APPENDIX B

MASTER-AID CLASSIFICATION ALGORITHM

USER'S MANUAL

MANUAL FOR MASTER-AID PROGRAM

by John Gins

1.0 INTRODUCTION

This manual describes the uses and access rules for two very similar programs that form decision trees. The first program, Master-Aid-Classification,performs non-parametric discriminant analysis on data sets that can be classified into two classes (an n-class version may be implemented in the future). The second program, Master-Aid-Splitting, performs the standard AID algorithm on continuous data sets.

Since the second program is still under development, the first draft of the manual will describe the first program only.

2.0 MASTER-AID-CLASSIFICATION PROGRAM

This program implements the Friedman algorithm, the weighted Friedman algorithm with the Breiman cost function, the Stone algorithm, and the Aid Classification algorithm. Since these are non-parametric statistics, no prior knowledge of the distribution of the various variables is necessary. The problem of unequal class sizes observed with the Friedman algorithm is handled by the other three algorithms. Section 2.1 describes the input and output of this program, and Section 2.2 describes the way the program works.

2.1 INPUT AND OUTPUT OF THE MASTER-AID-CLASSIFICATION PROGRAM

The program requires two general types of input files: data files and an information file. The information file is described in Section 2.1.1, the data file in Section 2.1.2, and the output in Section 2.1.3.

2.1.1 Information File

The first two records of the information file are in free format. All values should be separated by either a comma or a space. Alphabetic values should have single quotes around them. Values in exponential format should have no space between the fractional part and the exponent (i.e., 1.030E17).

The first record contains information about the data input files, the location of the format for a test file, and the number of variables in the data vector. Table 2.1.1 describes this record.

The second record contains further information about the data input files and the algorithms to be used. Printing information and decision

Table 2.1.1 First Information Record

VARIABLE NAME	TYPE	DESCRIPTION
INFILE	INTEGER	The FORTRAN number of the learning data file to be used. This should not be 6, 7, or 5. If this value is 0 then the learn mode will be skipped (refer to ITSAVE).
INTEST	INTEGER	The FORTRAN number of the testing data file to be used. This should not be 6, 7, 5 or the same as INFILE. If this value is 0 then the test mode will be skipped.
ISAME	INTEGER	This is a flag indicating if further information is to be read for the test mode. 1 means read new information; 0 means use old information.
LREC	INTEGER	This is the number of variables that can be used for splitting. $1 \le LREC \le 48$. The value of LREC directly affects the number of records that can be processed. The total number of data points that can be processed is approximately $\frac{80000}{2LREC+4}$.

rules are also specified. Table 2.1.2 describes this record. A few of the options are further clarified in the following paragraphs.

Only one number is allowed as the missing value indicator, VMIS; therefore, the input data sets will have to be internally consistent with respect to missing value indicators (i.e., if -9999.0 is used as a missing value for one variable, it should be used for all). Use of blanks for missing values is unwise because, when read in by FORTRAN, a blank is given a value zero (0). However, zero is a legal missing value.

The choice of algorithm (NALG) is important to the user of this program. The Friedman algorithm (NALG=1) was originally intended to handle data having a balanced design model (i.e., 50% of the data in class 1 and 50% in class 2). A balanced design model is not always the case, however. It has been suggested that, unless a balanced model can be assured, the weighted Friedman algorithm be used (NALG=0). Friedman and weighted Friedman algorithms are the same if the learning data set contains an equal number of vectors assigned to each class and RLAMDA=0. RLAMDA is a cost parameter used to weight splits that create pure nodes more heavily. The Stone splitting algorithm (NALG=2) and the Aid Classification algorithm (NALG=3) can also be accessed in the current version of the program.

The header vector, which can contain up to 100 real or A4 values, is used only when the record is read in the first time. The only important piece of information that must be in the header vector is the class label for the record. This value is converted to either a 1, for class 1, or a 2, for class 2.

Table 2.1.2 Second Information Record

VARIABLE NAME	TYPE	DESCRIPTION
CLASS1	REAL or ALPHA	The value of the class label (see INDCL) that corresponds to the first of two classes.
CLASS2	REAL or ALPHA	The value of the class label (see INDCL) that corresponds to the second of two classes. If a blank value (i.e., ' ') is input,then any class label not equal to CLASS1 will be assigned to the second of the two classes.
VMIS	REAL	This is the missing value code for all variables of the splitting vector.
ITSAVE	INTEGER	The FORTRAN unit number of the data set where the tree data should be saved if INFILE≠0. If INFILE=0,then tree data should be read from ITSAVE. If ITSAVE=0,then do not save the tree. ITSAVE≠7,6,5,INFILE, or INTEST.
NALG	INTEGER	This is a flag to indicate choice of algorithm.If NALG=1, use the unweighted Friedman algorithm for splitting.If NALG=0, the weighted Friedman algorithm is used.If NALG=2,the Stone algorithm is used. If NALG=3,the Aid Classification algorithm is used.
RLAMDA	REAL	The cost for incorrect classification using the weighted Friedman algorithm RLAMDA ≥ 0. If NALG≠0, RLAMDA should be 0.
NHEAD	INTEGER	The number of variables in the header of each data record (1≤NHEAD≤100).The class label must be contained in this label. None of the other information in the header vector is saved internally by the program.
INDCL	INTEGER	The location in the header of the class label of the record. (1≤INDCL≤NHEAD.)
NFORM	INTEGER	The number of data format cards to be read as information after this information record. (See information record 3 (1≤NFORM≤20).)
NRULE1	INTEGER	The smallest number of data points for which a node can be further split. If the number of data points at a node is less than NRULE 1,then stopping RULE 1 is invoked.
RULE2	REAL	Largest percentage(50% to 100%)allowed for the probability of correct classification. If the probability of correct classification is greater than RULE2,then stopping RULE 2 is invoked.

Table 2.1.2 Second Information Record (Cont'd)

VARIABLE NAME	TYPE	DESCRIPTION
RULE3	REAL	Smallest allowable Beta value of the distance term between 0.0001 and 1.0. This rule can be suppressed (see NODEC below).
NCOICE	INTEGER	Classification algorithm flag. NCOICE=1 means use majority rule to determine the classification of a node. NCOICE=2 means that the class with the largest probability of correct classification value is used to classify a node.
NODEC	INTEGER	This flag governs the type of nodes to be printed by the program. Use of this parameter actually can provide a limited amount of look-ahead capability. NODEC=0 Print all terminal nodes and those split nodes that are valid with respect to the four stopping rules. NODEC=1 Print all terminal nodes and all split nodes that are valid with respect to stopping rules 1, 2, and 4. If a node split violates rule 3 but has a descendant that is valid with respect to rule 3,then a split is allowed. NODEC=2 Print all terminal nodes and all split nodes that are valid with respect to stopping rules 1, 2, and 4.
NTREE	INTEGER	Is a flag which,if 1,writes data to FORTRAN unit 7,which is compatible with Dennis Bicker's tree-drawing program.
NCNAME	INTEGER	Is a flag which,if 1,means read in class names after the format cards. (Default names are CLASS1 and CLASS2.)
NVNAME	INTEGER	Is a flag which,if 1,means read in the LREC variable names after the class names.
NINDX	INTEGER	Is the total number of variables used in the classification process during this run. $1 \leq NINDX \leq LREC$.
INDXV	NINDX INTEGERS	These NINDX numbers are the vector pointers indicating the order in which the input variables should be processed at each node. $1 \leq INDXV(i) \leq LREC$ for all i such that $1 \leq i \leq NINDX$.

There are four termination rules used in this program. A termination is a decision not to split a node. Nodes which are not split are called terminal nodes and are assigned a class. The first rule, NRULE1, specifies the smallest number of values for which a split can be determined. NRULE1 is user-determined and is usually equal to the square root of the total number of cases. RULE2 stops further splitting by checking the probability of correct classification once the node is reached. If the probability is very high (i.e., 97.5%), then further splitting may not be necessary. RULE3, the Beta rule, stops further splitting if a split will not explain very much about the process (Beta is too small); it can be suspended either temporarily or permanently (see NODEC) for a given run. RULE4 sets an absolute bound on the minimum amount of explanation (Beta) required before splitting (RULE4=0.0001).

The array INDSV contains the indexes of the variables in each record that are to be processed and in what order they should be processed. The order is important only when ties occur in the parameters used for splitting. It a tie occurs, the first variable is used for splitting.

The third through sixth records are read in, using 10A8 format. These contain information on format and names. These cards are described in Table 2.1.3.

2.1.2 The Data Files

Two types of files can be used. The learning data file is on FORTRAN unit INFILE, if INFILE ≠0. This data is described with the format data on the information records. If the test data file (INTEST≠0, INTEST≠INFILE) has a different file structure, then ISAME=0 and another set of information cards 2 through 6 must be read in. Table 2.1.4 describes a typical record.

Table 2.1.3 Information Records 3 through 6

RECORD	VARIABLE NAME	TYPE	DESCRIPTION
3	AFORM	ALPHA	Contains NFORM*80 columns of valid FORTRAN format describing the data input.
4	CLNAME	ALPHA	Two names of 8 columns each. The first is the name given class 1, and the second name, class 2. (This is only used if NCNAME=1.)
5	VNAME	ALPHA	Contains the LREC names for each variable in input order. The name length is 8 columns and there are ten to a card. (This is only used if NVNAME=1.)
6	ATITLE	ALPHA	Contains an 80-column title. This is optional but must be at least a blank line if ISAME=1. The title is put on each output sheet and on the tree.

Table 2.1.4 Typical Data Record

VARIABLE NAME	TYPE	DESCRIPTION
HEAD	REAL or ALPHA	This is the header information read of the record. There are NHEAD values read in,and the INDCLth value is the class indicator of the record.
ALL	REAL	This is a real vector containing the LREC values read off the record. These LREC values are used to define the decision tree.

2.1.3 Output for the Master-Aid Classification Program

Three types of output records are generated from this program. Two are data files and one is printout.

If NTREE = 1, then a file containing node and split data is written on FORTRAN unit 7. This information is in a form that can be read by Dennis Bicker's tree plot program.

If ITSAVE is a valid FORTRAN unit number, then the matrix containing all node information is written. This data, described in Table 2.1.5, is used as input to the program if only the test mode is used.

Printout for this program is of three different types in learning mode and two different types in test mode, as well as a printout of input information used in both modes.

For each node examined in learning mode, a detailed breakdown of the values for distance, the cutpoint theta, and the Beta value for each variable are given.

A node summary is given which tabulates the contents of the array CUT. Part of the array description is given in Table 2.1.5, with the following exceptions:

- CUT(2) is the number of vectors in class 1 with respect to CUT(1)
- CUT(3) is the number of vectors in class 2 with respect to CUT(2)
- CUT(11) is the number of vectors in class 1 split from CUT(7)
- CUT(12) is the number of vectors in class 2 split from CUT(7)
- The variable name, distance, Beta, and cutpoint are left blank for terminal nodes.

Table 2.1.5 Data Tree Written--to (Read From) a Save File

RECORD	VARIABLE NAME	VARIABLE FORMAT	COLUMNS	DESCRIPTION
				Occurs once
1	Q1	F10.8	1-10	Probability of misclassifying class 1.
1	Q2	F10.8	11-20	Probability of misclassifying class 2.
				Repeated as many times as nodes
2	I	I5	1-5	Node number of this record.
2	CUT(1)	F10.0	6-15	The variable number for which this node was split.
2	CUT(2)	F10.0	16-25	The number of learning vectors in class 1 which descended from the previous node (CUT(7)).
2	CUT(3)	F10.0	26-35	The number of learning vectors in class 2 which descended from the previous node (CUT(7)).
2	CUT(4)	F10.5	36.45	Kolmogorov-Smirnov distance for the split with respect to CUT(1).
2	CUT(5)	F10.5	46-55	BETA value for CUT(4) (corresponds to the amount of explanation of the classification CUT(4) makes).
2	CUT(6)	F10.5	56-65	The splitting point for CUT(1): everything less than or equal to goes to the left descendant, greater than,to the right.
2	CUT(7)	F5.0	66-70	Node number this node descended from.
2	CUT(8)	F5.0	71-75	Node number of left descendant node.
2	CUT(9)	F5.0	76-80	Node number of right descendant node.
2	CUT(10)	F5.0	81-85	Type of descendant:0 means top node, -1 means left descendant of CUT(7), 1 means right descendant of CUT(8).
2	CUT(13)	F5.0	86-90	Stopping rule applied:1,2,3,4 or 0.

Table 2.1.5 Data Tree Written--to (Read From) a Save File (Cont'd)

RECORD	VARIABLE NAME	VARIABLE FORMAT	COLUMNS	DESCRIPTION
2	CUT(14)	F5.0	91-95	Class of node:1 = class 1, 2 = class 2.
2	CUT(15)	F5.0	96-100	0 if class was selected normally; 1. randomly.
2	CUT(16)	F5.0	101-105	0 if CUT(1) was unique "best";1, ambiguous.
2	CUT(17)	F5.0	106-110	% Probability of correct classification, given that this node was reached.

- "TERMINAL NODE" replaces descendant node numbers for terminal nodes
- Descendant type CUT(10) is replaced by "LEFT", " ", or "RIGHT" instead of -1, 0, 1, respectively
- Where possible,the class name is used instead of a number
- CUT(15) = 1 is replaced by "*", and CUT(16) = 1 is replaced by "#". Footnotes are given explaining these abnormal situations
- If any prior probabilities are used in classification, they are listed at the bottom of the page.

A misclassification summary is given on a separate page. This summary tabulates the number of correct and incorrect classifications made by the tree described above. The percentages of classifications correct and wrong are also given.

In the test mode, a set of data is run through a previously constructed decision tree. A node summary and a misclassification summary are given for the test. The printout is the same as for the learn mode,with the following exceptions:

- The vectors for classes 1 and 2 given in CUT (2) and CUT(3) are the same as those in CUT (11) and CUT (12) in the learning mode
- The number of vectors listed in CUT(11) and CUT(12) are the number of test vectors in classes 1 and 2, respectively, that were split from the previous node (CUT(7))
- All the rest of the information in this node summary was taken from the learn mode
- The misclassification summary reflects only the vectors that were classified during the test

- An estimate is made of the probability of correct and wrong classification,given the test set and probability of misclassification from the learn set.

2.2 PROGRAM DESCRIPTION

The next two sections briefly describe how the program works and its limitations. The weighted and unweighted Friedman algorithm will be more completely described in a paper by Leo Breiman.

2.2.1 Learning Mode

The learning mode is turned on when INFILE is given a valid unit number. The current limitations are as follows: $1 \leq \text{LREC} \leq 48$. The number of valid records used in learning, NT, must be greater than zero and $2\text{NT}(\text{LREC} + 2) \leq 80000$. The largest number of nodes that can be calculated is 1000.

The main program reads in the first information card from FORTRAN unit 5. It determines the dimensions of the array ALL,which is used to store and process the learning vectors.

The main subroutine MAINP is used to do most of the data processing and I/O. After the rest of the information is read in, the class of the record from HEAD(INDCL) is read and checked with the valid class codes. If it fails, then the next card is read;if it passes, then all LREC vector variables (independent variables) are placed in the array ALL.

Unless all the data is in one class, a split is always made on the first node. For all other vectors, termination rules 1 and 2 are applied before splitting is attempted. Rules 3 and 4 are applied after a splitting criterion has been determined. Classification of a terminal node is always made before splitting is attempted.

In general, a split (cut) point is determined as follows:

- The variables are treated individually
- They are sorted without missing values,and the class is kept with each value
- The maximum distance is determined for this variable, and the associated Beta and split point are saved. (The value of the variable at which the distance is maximum is the split point.)
- Notice that this method might result in each variable set using a different number of vectors since missing values are only taken into account during the sorting process. Thus,CUT(2) $\leq$ CUT(11) and CUT(3) $\leq$ CUT(12).

After a split is determined,and if all the stopping rules are satisfied, then the data set is split into two parts, left and right descendants, and these parts are moved to different parts of the array ALL. Processing continues after all terminal nodes are reached, and summary tables are printed.

2.2.2 Testing Mode

The test mode is turned on when INTEST is given a valid value. If INFILE=0, then a decision tree is read in from FORTRAN unit ITSAVE. Currently, the data is read and processed exactly as in the learning mode, except that no distance is calculated, since the split point from the learning mode is used at each node.

APPENDIX C

AS-51 LOG-LINEAR MODEL FIT ALGORITHM

Algorithm AS 51

Log-linear Fit for Contingency Tables

By S. J. HABERMAN

Department of Statistics, University of Chicago

LANGUAGE

ANSI Standard Fortran

DESCRIPTION AND PURPOSE

This algorithm performs an iterative proportional fit of the marginal totals of a contingency table. The method used has been described by Deming and Stephan (1942), Fienberg (1970) and Goodman (1970). The algorithm may be used to obtain maximum likelihood estimates which correspond to hierarchical log-linear models for both complete and incomplete contingency tables.

In hierarchical log-linear models for complete tables, the logarithms $\{\mu_{i_1 \ldots i_d}\}$ of the expected values $\{m_{i_1 \ldots i_d}\}$ of the frequencies $\{n_{i_1 \ldots i_d}\}$ of a factorial table with $1 \leqslant i_j \leqslant r_j$ for $j = 1, \ldots, d$ are assumed to satisfy an additive model similar to those encountered in analysis of variance. In such a model,

$$\mu_{i_1 \ldots i_d} = v^{\phi} + \sum_{k=1}^{l} v^{B_k}_{\{i_j:\, j \in B_k\}}, \tag{1}$$

where the non-empty sets B_k, $k = 1, \ldots, l$, are subsets of the sets of integers from 1 to d. The terms $v^{B_k}_{\{i_j:\, j \in B_k\}}$ satisfy the standard constraints. For example, if in a $2 \times 3 \times 4$ table,

$$\mu_{i_1 i_2 i_3} = v^{\phi} + v^{1}_{i_1} + v^{2}_{i_2} + v^{3}_{i_3} + v^{12}_{i_1 i_2} + v^{13}_{i_1 i_3} + v^{23}_{i_2 i_3}, \tag{2}$$

then

$$\sum_{i_1=1}^{2} v^{1}_{i_1} = \sum_{i_2=1}^{3} v^{2}_{i_2} = \sum_{i_3=1}^{4} v^{3}_{i_3} = 0$$

and

$$\begin{aligned} \sum_{i_1=1}^{2} v^{12}_{i_1 i_2} = \sum_{i_2=1}^{3} v^{12}_{i_1 i_2} = \sum_{i_1=1}^{2} v^{13}_{i_1 i_3} = \sum_{i_3=1}^{4} v^{13}_{i_1 i_3} \\ = \sum_{i_2=1}^{3} v^{23}_{i_2 i_3} = \sum_{i_3=1}^{4} v^{23}_{i_2 i_3} = 0. \end{aligned} \tag{3}$$

In this model, $d = 3$, $B_1 = \{1\}$, $B_2 = \{2\}$, $B_3 = \{3\}$, $B_4 = \{1, 2\}$, $B_5 = \{1, 3\}$ and $B_6 = \{2, 3\}$. In a hierarchical model, it is assumed that if C is a non-empty subset of B_k for some k, $1 \leqslant k \leqslant l$, then $C = B_j$ for some j, $1 \leqslant j \leqslant l$. Thus the example defined by (2) is a hierarchical model.

In any hierarchical model, sets C_k, $k_j = 1, \ldots, s$ may be found such that each C_k is equal to some $B_{k'}$, each B_k is contained in some $C_{k'}$, and no C_k is contained in any other set $C_{k'}$, $1 \leqslant k' \leqslant s$. The sets $C_1 = \{1, 2\}$, $C_2 = \{1, 3\}$ and $C_3 = \{2, 3\}$ satisfy these requirements in the example of the $2 \times 3 \times 4$ table. If the frequencies in the table are generated from independent Poisson samples or from a multinomial sample, the maximum likelihood estimate $\{\hat{m}_{i_1\cdots i_d}\}$ of $\{m_{i_1\cdots i_d}\}$ has the same marginal totals $\hat{m}^{C_k}_{\{i_j:\, j\in C_k\}}$ as $\{n_{i_1\cdots i_d}\}$ for $k = 1, \ldots, r$. In the example,

$$\hat{m}^{12}_{i_1 i_2} = \hat{m}_{i_1 i_2 +} = n_{i_1 i_2 +} = n^{12}_{i_1 i_2}, \tag{4}$$

$$\hat{m}^{13}_{i_1 i_3} = \hat{m}_{i_1 + i_3} = n_{i_1 + i_3} = n^{13}_{i_1 i_3} \tag{5}$$

and

$$\hat{m}^{23}_{i_2 i_3} = \hat{m}_{+ i_2 i_3} = n_{+ i_2 i_3} = n^{23}_{i_2 i_3}. \tag{6}$$

Thus a hierarchical model may be described in terms of the marginal totals to be fit.

The Deming–Stephan algorithm is an iterative method which finds $\{\hat{m}_{i_1\cdots i_d}\}$. The basic iteration cycle begins with an initial estimate $\{m^{(0)}_{i_1\ldots i_d}\}$ of $\{\hat{m}_{i_1\cdots i_d}\}$ such that $\{\log m^{(0)}_{i_1\ldots i_d}\}$ satisfies (1). At the first cycle, one may set each $m_{i_1\cdots i_d} = 1$ for $1 \leqslant i_j \leqslant r_j$ and $j = 1, \ldots, d$. An iteration cycle consists of s steps. Each step is defined by the equation

$$m^{(k)}_{i_1\ldots i_d} = m^{(k-1)}_{i_1\ldots i_d} \frac{n^{C_k}_{\{i_j:\, j\in C_k\}}}{[m^{(k-1)}]^{C_k}_{\{i_j:\, j\in C_k\}}}. \tag{7}$$

At the end of a cycle, one may let $m^{(0)}_{i_1\ldots i_d}$ for the next cycle be $m^{(s)}_{i_1\ldots i_d}$. In the example,

$$m^{(1)}_{i_1 i_2 i_3} = m^{(0)}_{i_1 i_2 i_3} \frac{n_{i_1 i_2 +}}{m^{(0)}_{i_1 i_2 +}}, \tag{8}$$

$$m^{(2)}_{i_1 i_2 i_3} = m^{(1)}_{i_1 i_2 i_3} \frac{n_{i_1 + i_3}}{m^{(1)}_{i_1 + i_3}} \tag{9}$$

and

$$m^{(3)}_{i_1 i_2 i_3} = m^{(2)}_{i_1 i_2 i_3} \frac{n_{+ i_2 i_3}}{m^{(2)}_{+ i_2 i_3}}. \tag{10}$$

The algorithm may also be applied to incomplete tables in which

$$n_{i_1 i_2 \cdots i_d} = m_{i_1 \cdots i_d} = 0$$

for certain indices and (1) holds for the remaining indices. In such tables, it is sufficient to begin the initial cycle with $m^{(0)}_{i_1\ldots i_d} = 0$ when $m_{i_1\ldots i_d} = 0$ and $m^{(0)}_{i_1\ldots i_d} = 1$ when $m_{i_1\ldots i_d} > 0$. Deming and Stephan (1942) use this algorithm to adjust tables so that they have specified marginal totals.

STRUCTURE

SUBROUTINE LOGLIN (NVAR, DIM, NCON, CONFIG, NTAB, TABLE, FIT, LOCMAR, NMAR, MARG, NU, U, MAXDEV, MAXIT, DEV, NLAST, IFAULT)

Formal parameters

NVAR	Integer	input:	the number of variables d in the table.
DIM	Integer array (*NVAR*)	input:	the number of categories r_j in each variable of the table.
NCON	Integer	input:	the number s of marginal totals to be fit.
CONFIG	Integer array (*NVAR, NCON*)	input:	the sets C_k, $k = 1, \ldots, s$, indicating marginal totals to be fit.

Reproduced from
best available copy.

NTAB	Integer	input:	the number of elements in the table.
TABLE	Real array (*NTAB*)	input:	the table to be fit.
FIT	Real array (*NTAB*)	input and output:	the fitted table.
LOCMAR	Integer array (*NCON*)	output:	pointers to the tables in *MARG*.
NMAR	Integer	input:	the dimension of *MARG*.
MARG	Real array (*NMAR*)	output:	the marginal tables to be fit.
NU	Integer	input:	the dimension of *U*.
U	Real array (*NU*)	output:	a work area used to store fitted marginal tables.
MAXDEV	Real	input:	the maximum permissible difference between an observed and fitted marginal total.
MAXIT	Integer	input:	the maximum permissible number of iterations.
DEV	Real array (*MAXIT*)	output:	*DEV*(*I*) is the maximum observed difference encountered in iteration cycle *I* between an observed and fitted marginal total.
NLAST	Integer	output:	the number of the last iteration.
IFAULT	Integer	output:	an error indicator. See *failure indications* below.

Comments

The tables *TABLE* and *FIT* are arranged in standard Fortran fashion. If the observations to be fit are represented in an array $\{n_{i_1 i_2 i_3}\}$, where $1 \leqslant i_1 \leqslant 2$, $1 \leqslant i_2 \leqslant 3$ and $1 \leqslant i_3 \leqslant 4$, then $NVAR = 3$, $DIM(1) = 2$, $DIM(2) = 3$, $DIM(3) = 4$, $NTAB = 24$, $TABLE(1) = n_{111}$, $TABLE(2) = n_{211}$, $TABLE(3) = n_{121}$, $TABLE(4) = n_{221}$, etc. If the fitted table is an array $\{\hat{m}_{ijk}\}$, then $FIT(1) = \hat{m}_{111}$, $FIT(2) = \hat{m}_{211}$, $FIT(3) = \hat{m}_{121}$, etc.

Marginal totals to be fitted are described by the array *CONFIG*. If C_k has d_k elements, then these elements are given by $CONFIG(J, K)$, where $1 \leqslant J \leqslant d_k$. If $d_k < d$, $CONFIG(J, K) = 0$ for $d_k < J \leqslant d = NVAR$. If maximum likelihood estimates for the table in the preceding paragraph are found under the assumption that (2) holds, then the marginal totals $n_{i_1 i_2 +}$, $n_{i_1 + i_3}$ and $n_{+ i_2 i_3}$ corresponding to $\{1, 2\}$, $\{1, 3\}$ and $\{2, 3\}$ are fitted. Thus $NCON = 3$, $CONFIG(1, 1) = 1$, $CONFIG(2, 1) = 2$, $CONFIG(3, 1) = 0$, $CONFIG(1, 2) = 1$, etc.

FIT is initially set as an estimate of the fitted table. If a complete table is fit to a log-linear model corresponding to the given marginals, one may let $FIT(I) = 1$ for $1 \leqslant I \leqslant NTAB$. If an incomplete table is fit, then one may let $FIT(I) = 1$ if the corresponding fitted value can be positive and $FIT(I) = 0$ if the corresponding fitted value must be zero. If in the three-way table under discussion, n_{111} cannot be observed, one would set $FIT(1) = 0$ and $FIT(I) = 1$ for $2 \leqslant I \leqslant NTAB$.

An iterative procedure is used to transform *FIT* into the desired fitted table. As many as *MAXIT* iterations may be performed, where *MAXIT* may be set as 10 or 15 in most problems. Iteration *I* is the last iteration if $DEV(I) < MAXDEV$. A reasonable value for *MAXDEV* is often 0·25.

Reproduced from
best available copy.

Under some circumstances, convergence occurs after one cycle. In such cases, *MAXIT* may be set to 1 to avoid an unnecessary second cycle. No error condition for failure of convergence will be observed in this case.

The marginals to be fit are placed in *MARG* by the subroutine. They are arranged in the order specified by *CONFIG*. In the example under discussion, *NMAR* must be at least 26. The array *MARG* is constructed so that $MARG(1) = n_{11+}$, $MARG(2) = n_{21+}$, $MARG(3) = n_{12+}, \ldots, MARG(7) = n_{1+1}$, $MARG(8) = n_{2+1}$, etc. To facilitate reference to marginal tables, locations of first elements of marginal tables are stored in *LOCMAR*. Thus $LOCMAR(1) = 1$, $LOCMAR(2) = 7$ and $LOCMAR(3) = 15$.

During computations, a work area *U* is required. This array should be as large as the largest marginal table used.

Failure indications

If *IFAULT* = 0, then no error was detected.

If *IFAULT* = 1, then *CONFIG* contains an error.

If *IFAULT* = 2, then *MARG* or *U* is too small.

If *IFAULT* = 3, then the algorithm did not converge to the desired accuracy within *MAXIT* iterations.

If *IFAULT* = 4, then *DIM, TABLE, FIT* or *NVAR* is erroneously specified.

RESTRICTIONS

The number of variables *NVAR* must not exceed 7, unless alterations are made as specified in the algorithms.

ACKNOWLEDGEMENT

This algorithm was written as part of research sponsored in part by National Science Foundation research grant No. NSF GS 2818.

REFERENCES

DEMING, M. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.*, 11, 427–~~444~~.

FIENBERG, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.*, 41, 907–917.

GOODMAN, L. A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Statist. Ass.* 65, 226–256.

```
      SUBROUTINE LOGLIN(NVAR,DIM,NCON,CONFIG,NTAB,TABLE,FIT,LOCMAR,
     *  NMAR,MARG,NU,U,MAXDEV,MAXIT,DEV,NLAST,IFAULT)
C
C        ALGORITHM AS 51  APPL.STATIST. (1972) VOL.21, NO.2
C
C        PERFORMS AN ITERATIVE PROPORTIONAL FIT OF THE MARGINAL TOTALS
C        OF A CONTINGENCY TABLE.
C
C        THIS VERSION PERMITS UP TO SEVEN VARIABLES. IF THIS LIMIT IS
C        TOO SMALL, CHANGE THE VALUE OF MAXVAR, AND OF THE DIMENSION
C        IN THE DECLARATION OF CHECK AND ICON - SEE ALSO THE CHANGES
C        NEEDED IN AS 51.1 AND AS 51.2
C
      REAL MARG(NMAR),U(NU),TABLE(NTAB),FIT(NTAB),MAXDEV,DEV(MAXIT)
      INTEGER DIM(NVAR),CHECK(7),CONFIG(NVAR,NCON),LOCMAR(NCON),
     *  POINT,SIZE,ICON(7)
C
      DATA MAXVAR/7/
```

Reproduced from
best available copy.

```
C
C        CHECK VALIDITY ØF NVAR, THE NUMBER ØF VARIABLES,
C        AND ØF MAXIT, THE MAXIMUM NUMBER ØF ITERATIØNS
C
      IF(NVAR.GT.0.AND.NVAR.LE.MAXVAR.AND.MAXIT.GT.0)GØTØ10
    5 IFAULT=4
      RETURN
C
C        LØØK AT TABLE AND FIT CØNSTANTS
C
   10 SIZE=1
      DØ 30 J=1,NVAR
      IF(DIM(J).LE.0)GØTØ5
   20 SIZE=SIZE*DIM(J)
   30 CØNTINUE
      IF(SIZE.LE.NTAB)GØTØ40
   35 IFAULT=2
      RETURN
   40 X=0.0
      Y=0.0
      DØ 60 I=1,SIZE
      IF(TABLE(I).LT.0.0.ØR.FIT(I).LT.0.0)GØTØ5
      X=X+TABLE(I)
      Y=Y+FIT(I)
   60 CØNTINUE
C
C         MAKE A PRELIMINARY ADJUSTMENT TØ ØBTAIN THE FIT TØ AN
C         EMPTY CØNFIGURATIØN LIST
C
      IF(Y.EQ.0.0)GØTØ5
      X=X/Y
      DØ 80 I=1,SIZE
   80 FIT(I)=X*FIT(I)
C
C        ALLØCATE MARGINAL TABLES
C
      PØINT=1
      DØ 150 I=1,NCØN
C
C        A ZERØ BEGINNING A CØNFIGURATIØN INDICATES THAT THE LIST IS
C        CØMPLETED
C
      IF(CØNFIG(1,I).EQ.0)GØTØ160
C
C        GET MARGINAL TABLE SIZE. WHILE DØING THIS TASK, SEE IF THE
C        CØNFIGURATIØN LIST CØNTAINS DUPLICATIØNS ØR ELEMENTS ØUT ØF
C        RANGE
C
      SIZE=1
      DØ 90 J=1,NVAR
   90 CHECK(J)=0
      DØ 120 J=1,NVAR
      K=CØNFIG(J,I)
C
C        A ZERØ INDICATES THE END ØF THE STRING
C
      IF(K.EQ.0)GØTØ130
C
C        SEE IF ELEMENT VALID
C
      IF(K.GE.0.AND.K.LE.NVAR)GØTØ100
   95 IFAULT=1
      RETURN
C
C        CHECK FØR DUPLICATIØN
C
  100 IF(CHECK(K).EQ.1)GØTØ95
  110 CHECK(K)=1
```

```
C
C       GET SIZE
C
    SIZE=SIZE*DIM(K)
 120 CØNTINUE
C
C       SINCE U IS USED TØ STØRE FITTED MARGINALS, SIZE MUST NØT
C       EXCEED NU
C
 130 IF(SIZE.GT.NU)GØTØ35
C
C       LØCMAR PØINTS TØ MARGINAL TABLES TØ BE PLACED IN MARG
C
 140 LØCMAR(I)=PØINT
    PØINT=PØINT+SIZE
 150 CØNTINUE
C
C       GET N, NUMBER ØF VALID CØNFIGURATIØNS
C
    I=NCØN+1
 160 N=I-1
C
C       SEE IF MARG CAN HØLD ALL MARGINAL TABLES
C
    IF(PØINT.GT.NMAR+1)GØTØ35
C
C       ØBTAIN MARGINAL TABLES
C
 170 DØ 190 I=1,N
    DØ 180 J=1,NVAR
 180 ICØN(J)=CØNFIG(J,I)
    CALL CØLLAP(NVAR,TABLE,MARG,1,LØCMAR(I),NTAB,NMAR,DIM,ICØN,.TRUE.)
 190 CØNTINUE
C
C       PERFØRM ITERATIØNS
C
    DØ 220 K=1,MAXIT
C
C       XMAX IS MAXIMUM DEVIATIØN ØBSERVED BETWEEN FITTED AND TRUE
C       MARGINAL DURING A CYCLE
C
    XMAX=0.0
    DØ 210 I=1,N
    DØ 200 J=1,NVAR
 200 ICØN(J)=CØNFIG(J,I)
    CALL CØLLAP(NVAR,FIT,U,1,1,NTAB,NU,DIM,ICØN,.TRUE.)
    CALL ADJUST(NVAR,FIT,U,MARG,1,1,LØCMAR(I),NTAB,NU,NMAR,DIM,
   *  ICØN,XMAX)
 210 CØNTINUE
C
C       TEST CØNVERGENCE
C
    DEV(K)=XMAX
    IF(XMAX.LT.MAXDEV)GØTØ240
 220 CØNTINUE
    IF(MAXIT.GT.1)GØTØ230
    NLAST=1
    RETURN
C
C       NØ CØNVERGENCE
C
 230 IFAULT=3
    NLAST=MAXIT
    RETURN
C
C       NØRMAL TERMINATIØN
C
 240 NLAST=K
    RETURN
    END
```

Reproduced from
best available copy.

```
      SUBROUTINE COLLAP(NVAR,X,Y,LOCX,LOCY,NX,NY,DIM,CONFIG,OPTION)
C
C        ALGORITHM AS 51.1 APPL.STATIST. (1972), VOL.21, NO.2
C
C        IF OPTION IS TRUE, COMPUTES A MARGINAL TABLE FROM A COMPLETE
C        TABLE. IF OPTION IS FALSE, REMOVES AN EFFECT FROM A TABLE.
C        ALL PARAMETERS ARE ASSUMED VALID - THERE IS NO TEST FOR ERRORS
C        IN INPUT.
C        IF THE VALUE OF NVAR IS TO BE GREATER THAN 7, THE DIMENSIONS IN
C        THE DECLARATIONS OF SIZE AND COORD MUST BE INCREASED TO THE
C        VALUES OF NVAR+1 AND NVAR, RESPECTIVELY
C
      INTEGER SIZE(8),DIM(NVAR),CONFIG(NVAR),COORD(7)
      LOGICAL OPTION
C
C        THE LARGER TABLE IS X AND THE SMALLER ONE IS Y
C
      REAL X(NX),Y(NY)
C
C        INITIALISE ARRAYS
C
      SIZE(1)=1
      DO 10 K=1,NVAR
      L=CONFIG(K)
      IF(L.EQ.0)GOTO20
      SIZE(K+1)=SIZE(K)*DIM(L)
   10 CONTINUE
C
C        FIND NUMBER OF VARIABLES IN CONFIGURATION
C
      K=NVAR+1
   20 N=K-1
C
C        IF MARGINAL TABLE DESIRED, INITIALISE Y. FIRST CELL OF
C        MARGINAL TABLE IS AT Y(LOCY) AND TABLE HAS SIZE(K) ELEMENTS
C
      IF(.NOT.OPTION)GOTO40
      LOCU=LOCY+SIZE(K)-1
      DO 30 J=LOCY,LOCU
   30 Y(J)=0.0
C
C        INITIALISE COORDINATES
C
   40 DO 50 K=1,NVAR
   50 COORD(K)=0
C
C        FIND LOCATIONS IN TABLES
C
      I=LOCX
   60 J=LOCY
      DO 70 K=1,N
      L=CONFIG(K)
      J=J+COORD(L)*SIZE(K)
   70 CONTINUE
      IF(OPTION)Y(J)=Y(J)+X(I)
      IF(.NOT.OPTION)X(I)=X(I)-Y(J)
C
C        UPDATE COORDINATES
C
      I=I+1
      DO 80 K=1,NVAR
      COORD(K)=COORD(K)+1
      IF(COORD(K).LT.DIM(K))GOTO60
      COORD(K)=0
   80 CONTINUE
      RETURN
      END
```

Reproduced from
best available copy.

```
      SUBRØUTINE ADJUST(NVAR,X,Y,Z,LØCX,LØCY,LØCZ,NX,NY,NZ,DIM,CØNFIG,D)
C
C        ALGØRITHM AS 51.2 APPL.STATIST. (1972), VØL.21, NØ.2
C
C        MAKES PRØPØRTIØNAL ADJUSTMENT CØRRESPØNDING TØ CØNFIG. ALL
C        PARAMETERS ARE ASSUMED VALID - NØ TEST IS MADE FØR ERRØRS.
C        IF THE VALUE ØF NVAR IS TØ BE GREATER THAN 7,THE DIMENSIØNS IN
C        THE DECLARATIØNS ØF SIZE AND CØØRD MUST BE INCREASED TØ THE
C        VALUES ØF NVAR+1 AND NVAR, RESPECTIVELY
C
      INTEGER SIZE(8),DIM(NVAR),CØNFIG(NVAR),CØØRD(7)
      REAL X(NX),Y(NY),Z(NZ)
C
C        SET SIZE ARRAY
C
      SIZE(1)=1
      DØ 10 K=1,NVAR
      L=CØNFIG(K)
      IF(L.EQ.0)GØTØ20
      SIZE(K+1)=SIZE(K)*DIM(L)
   10 CØNTINUE
C
C        FIND NUMBER ØF VARIABLES IN CØNFIGURATIØN
C
      K=NVAR+1
   20 N=K-1
C
C        TEST SIZE ØF DEVIATIØN
C
      L=SIZE(K)
      J=LØCY
      K=LØCZ
      DØ 30 I=1,L
      E=ABS(Z(K)-Y(J))
      IF(E.GT.D)D=E
      J=J+1
      K=K+1
   30 CØNTINUE
C
C        INITIALIZE CØØRDINATES
C
      DØ 40 K=1,NVAR
   40 CØØRD(K)=0
      I=LØCX
C
C        PERFØRM ADJUSTMENT
C
   50 J=0
      DØ 60 K=1,N
      L=CØNFIG(K)
      J=J+CØØRD(L)*SIZE(K)
   60 CØNTINUE
      K=J+LØCZ
      J=J+LØCY
C
C        NØTE THAT Y(J) SHØULD BE NØN-NEGATIVE
C
      IF(Y(J).LE.0.0)X(I)=0.0
      IF(Y(J).GT.0.0)X(I)=X(I)*Z(K)/Y(J)
C
C        UPDATE CØØRDINATES
C
      I=I+1
      DØ 70 K=1,NVAR
      CØØRD(K)=CØØRD(K)+1
      IF(CØØRD(K).LT.DIM(K))GØTØ50
      CØØRD(K)=0
   70 CØNTINUE
      RETURN
      END
```

Reproduced from
best available copy.

REFERENCES

Aerospace Corporation, Mobile Systems Group, 1976. "Short Test Correlation Analyses on 300, 1975 Model Year Cars, Volumes I and II." Los Angeles, California. (Prepared for Contract No. 68-01-0417, U.S. Environmental Protection Agency, Office of Air and Waste Management, Ann Arbor, Michigan, Publ. Nos. EPA-460/3-76-010a and EPA-460/3-76-010b.)

Afifi, A. A., and S. P. Azen, 1974. Statistical Analysis: A Computer Oriented Approach." Academic Press, Inc.

Anderson, T. W., 1958. An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Inc.

"Biomedical Computer Programs (BMDP)," 1975. University of California Press.

Bishop, Y. M., S. E. Feinberg, and P. W. Holland, 1975. Discrete Multivariate Analysis Theory and Practice." MIT Press.

Fox, D. J., and K. E. Guire, 1976. "MIDAS." Statistical Research Laboratory, University of Michigan (3rd edition).

Friedman, J. H., 1977. "A Recursive Partitioning Decision Rule for Nonparametric Classification." IEEE Trans. Computers, Vol. C-26, pp. 404-408.

Gnanadesikan, R., Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons.

Grizzle, J. E., C. F. Starmer, and Gary G. Koch, 1969. "Analysis of Categorical Data by Linear Models." Biometrics.

Haberman, S. J., 1972. "Log-linear Fit for Contingency Tables (Algorithm AS-51)." Appl. Statist., 21, pp. 218-225.

Hamilton Test Systems, 1977. "Plan of Performance for a Short Test Correlation and Effectiveness Study to the United States Environmental Protection Agency." (Prepared for Contract No. 68-03-2513, U.S. Environmental Protection Agency.)

Lachenbruch, P. A. and Mickey, M. R., 1968. "Estimation of Error Rates in Discriminant Analysis." Technometrics, 10, pp. 1-11.

McLachlan, G. J., 1975. "Confidence Intervals for the Conditional Probability of Misclassification in Discriminant Analysis." Biometrics, 31, pp. 161-167.

Okamoto, M., 1963. "An Asymptotic Expansion for the Distribution of the Linear Discriminant Analysis." Biometrics, pp. 1286-1293.

Sonquist, J. A., E. L. Baker, and J. N. Morgan, 1971. "Searching for Structure." Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

REFERENCES (Cont'd)

"Statistical Package for the Social Sciences (SPSS)," 1977. McGraw-Hill (2nd edition).